# MCESA CONTENT SPECIFIC ASSESSMENTS

## Technical Manual

MCESA
Maricopa County Education Service Agency

4041 N. Central Avenue, Suite 1200
Phoenix, AZ 85012
Phone: 602.506.3866 • Fax: 602.506.3753

Maricopa County Education Service Agency (MCESA), under the direction of Maricopa County Superintendent of Schools, is dedicated to ensuring that all school-age children in the county graduate college- and career-ready. MCESA builds alliance partnerships that provide leadership, services and programs in the areas of Educational Innovation, Economic Management and Executive Leadership.

Contact: laura.harnish@mcesa.maricopa.gov | 602-506-3550

## Contents

## Tables and Figures

# Part 1. Introduction

## 1.1 Rationale for Assessment Development

In 2011, new legislation, Arizona Revised Statute 15-203, required evaluation of educator effectiveness with measures of academic growth. At that time, the Arizona Instrument to Measure Standards (AIMS) was available to evaluate growth for reading, writing, and arithmetic. Teachers of these content areas were considered "Group A" teachers according to the Framework for Educator Effectiveness, having valid and reliable assessment data rot inform their evaluation score. However, a number of teachers in Arizona's schools teach subjects that fall outside the content that the AIMS tests measured. As a result, it was necessary that districts develop tests for "Group B" teachers, or those teachers who did not have valid and reliable assessments that could measure the progress or growth of students in non-AIMS tested content areas. MCESA began to develop assessments for non-traditionally tested subject areas including: social studies, physical education, art, dance, theater, and music.

## 1.2 Staged Development

MCESA planned to create pre-post pairs of assessments, called the MCESA Content Specific Assessments for 18 different courses. Creating this many new assessments required a staged approach and coordination of partners. Table 1 shows the three major stages of Development:

**Table 1 – Stages of Development**

| Stage | Subjects | Number of Courses |
|---|---|---|
| I | Fine Arts and PE assessments for elementary grades | 6 |
| II | Fine Arts and PE assessments for high school grades | 5 |
| III | Social Studies assessments for middle and high school grades | 7 |

## 1.3 Partners

MCESA engaged in numerous partnerships to create the MCESA Content Specific Assessments. The very first partnership created was the Cross-District Assessment Advisory Council. This council included administrators for research and assessment from county school districts as well as nationally known consultants in the field of assessment. This council helped to determine the design and format of the first assessments, which included 50-item, multiple-choice pre- and post-assessments for 18 courses and accompanying performance assessments for 12 of those courses. MCESA also partnered with WestEd, an agency vendor with recognized assessment development expertise, to design and employ a research-based process to create assessments. The partnership with WestEd generated immediate credibility for MCESA with county LEAs and yielded confidence in the end products. Throughout the development phases of the project, MCESA partnered with 45 school districts that provided 404 teachers to author and review items. Additionally, 20 different school districts, including those in the MCESA Rewarding Excellence in Instruction and Leadership (REIL) teacher incentive fund grant alliance, provided nearly 45,000 students to field test the assessment items. Finally, MCESA partnered with American Institutes for Research (AIR) to complete all psychometric processes associated with the assessments.

## Part 2. Involvement of Maricopa County Educators at All Levels

Maricopa County educators were integrally involved in test development for the pre- and post-assessments for Phase I, Phase II, and Phase III. Phase I included grades 3 and 8 physical education, grade 3 music, grades 3 and 8 visual arts, and choir. Phase II included band, dance, theater, high school physical education, and high school visual arts. Phase III included grades 6, 7, and 8 social studies, U.S. history, world history, economics, and civics/government.

Committees comprised of Maricopa County educators met throughout 2012 and 2013 in preparation for the 2013 and 2014 pre- and post-assessment field tests and the 2014 pre- and post-assessment operational tests. The educators serving on committees included teachers, curriculum specialists, and administrators who represented their respective content areas and grade levels. A bias/sensitivity committee was also convened to review items for any bias and sensitivity issues. The bias/sensitivity committee included a diverse group of community members.

MCESA conducted deliberate recruitment of facilitators and committee members for all events. For the content experts, MCESA specifically tried to balance the representatives across a variety of school districts representing the different geographic, economic, or political strata of the county. The bias review committee members were purposely recruited to represent different racial or ethnic populations as well as different economic regions of the county. All facilitators were trained in the assessment development process and best practices.

Committee meetings focused on the development of blueprints for each assessment, the development of item specifications for each assessment, and the development and review of all assessment items to be housed in the MCESA assessment item banks.

**Table 2 – Assessment Committee Meetings**

| Assessment Committee Meetings | | | |
|---|---|---|---|
| **Date** | **Phase** | **Assessment Type** | **Topic** |
| February 1–3, 2012 | I & II | Post | Blueprint and item specifications development |
| February 14–16, 2012 | I | Post | Item writing |
| March 21–23, 2012 | II | Post | Item writing |
| March 27, 2012 | I | Post | Bias review |
| March 28–29, 2012 | I | Post | Item content review |
| April 3-5, 2012 | III | Post | Blueprint and item specifications development |
| April 25-27, 2012 | III | Post | Item writing |
| May 2, 2012 | II | Post | Bias review |
| May 3–4, 2012 | II | Post | Item content review |
| May 31–June 1, 2012 | I & II | Pre | Blueprint and item specifications development |
| July 25, 2012 | I & II | Pre | Bias review |
| July 26–27, 2012 | I & II | Pre | Item content review |
| November 27–28, 2012 | III | Pre | Writing blueprints and item specifications finalization and item writing |
| January 30–31, 2013 | III | Pre | Blueprint and item specifications development |
| February 19, 2013 | III | Post | Bias review |
| February 20–21, 2013 | III | Post | Item content review |
| March 27–29, 2013 | III | Pre | Item writing |
| June 25, 2013 | III | Pre | Bias review |
| June 26, 2013 | III | Pre | Item content review |

The test development committee meetings included blueprint and item specifications development, during which Maricopa County educators reviewed the Arizona Performance Objectives (POs) and selected the priority POs to be assessed.

The test development meetings also included item writing workshops, in which MCESA educators wrote multiple-choice items aligned to the Arizona POs that were selected by the blueprint and item specifications development committee. The numbers of items written were based on item orders developed by WestEd, which were based on the blueprints and item specifications for each grade and subject.

The test development meetings also included bias/sensitivity reviews, in which committee members reviewed the items. Participants were recruited using the following process: MCESA contacted school districts and community organizations to recruit participants who represented the county in terms of geography, socio-economic background, race, and ethnicity. Racial or ethnic backgrounds included: Hispanic, Caucasian, African-American, and Native American participants. A WestEd and/or MCESA facilitator trained the committee on issues of bias and sensitivity.

The last type of test development meetings were item content reviews. In these meetings, 5–10 educators per grade and subject reviewed the items. Educators were recruited using the following process: school districts representing the range of the county were asked to identify their subject

matter leaders to participate in the development work. Each course had a balance of teachers representing a variety of demographic and socio-economic backgrounds.

All of the educators were trained by WestEd or MCESA facilitators (one facilitator per grade and subject) in group meetings at the item content review. Phase I and Phase II committee meetings were facilitated by WestEd, and Phase III meetings were facilitated by MCESA with WestEd support.

# Part 3. Test Design

## 3.1 Construct Measured: Content Standards for the Pre- and Post-Assessments

The MCESA Content Specific Assessments are designed to measure performance on the Arizona content standards adopted in 2009 for physical education; adopted in 2006 for visual arts, music, band, choir, theater, and dance; and adopted in 2005 and updated in 2006 for grades 6–8 social studies, U.S. history, world history, economics, and civics/government. These standards are organized by strand, concept, and performance objective. The pre- and post-assessments are based on the strands and concepts of the standards.

The following subjects are presented in the following tables:

- Physical education—grades 3, 8, and high school: Tables 3 -5
- Visual arts—grades 3, 8, and high school: Tables 6 - 8
- Music—grade 3: Table 9
- Band: Table 10
- Choir: Table 11
- Theater: Table 12
- Dance: Table 13
- Social studies—grades 6, 7, and 8: Table 14 - 16
- U.S. history: Table 17
- World history: Table 18
- Economics: Figure Table 19
- Civics/government: Table 20

Note: Not all Arizona strands and concepts are listed in the following tables. Some strands and concepts were deemed "performance-based assessment only" by the blueprint and item specifications development committee; therefore, they were not assessed in the multiple-choice pre- and post-assessments.

**Table 3 – Arizona Grade 3 Physical Education Concepts and Strands**

| Arizona Grade 3 Physical Education Concepts and Strands | |
|---|---|
| Strand 1. | Demonstrates competency in motor skills and movement patterns needed to perform a variety of physical activities |
| | Concept 1. Fundamental Movement Skills |
| Strand 2. | Demonstrates understanding of movement concepts, principles, strategies, and tactics as they apply to the learning and performance of physical activities |
| | Concept 1. Movement Concepts |
| | Concept 2. Scientific Principles |
| | Concept 3. Strategies and Tactics |
| Strand 4. | Achieves and maintains a health-enhancing level of physical fitness |
| | Concept 1. Health-Related Fitness |
| Strand 5. | Exhibits responsible personal and social behavior that respects self and others in physical activity settings |
| | Concept 1. Personal Behavior |
| | Concept 2. Social Behavior |

**Table 4 – Arizona Grade 8 Physical Education Concepts and Strands**

| Arizona Grade 8 Physical Education Concepts and Strands | |
|---|---|
| Strand 1. | Demonstrates competency in motor skills and movement patterns needed to perform a variety of physical activities |
| | Concept 1: Fundamental Movement Skills |
| | Concept 3: Complex or Specialized Movement Skills |
| Strand 2. | Demonstrates understanding of movement concepts, principles, strategies, and tactics as they apply to the learning and performance of physical activities |
| | Concept 1: Movement Concepts |
| | Concept 2: Scientific Principles |
| | Concept 3: Strategies and Tactics |
| Strand 3. | Participates regularly in physical activity both during and beyond the structured physical education class |
| | Concept 2: Physical Activity Outside a Physical Education Program |
| Strand 4. | Achieves and maintains a health-enhancing level of physical fitness |
| | Concept 1: Health-Related Fitness |
| Strand 5. | Exhibits responsible personal and social behavior that respects self and others in physical activity settings |
| | Concept 1: Personal Behavior |
| | Concept 2: Social Behavior |
| Strand 6. | Values physical activity for health, enjoyment, challenge, self-expression, and/or social interaction |
| | Concept 1: Values Physical Activity |

**Table 5 – Arizona High School Physical Education Concepts and Strands**

| Arizona High School Physical Education Concepts and Strands | |
|---|---|
| Strand 1. | Demonstrates competency in motor skills and movement patterns needed to perform a variety of physical activities |
| | Concept 3: Complex or Specialized Movement Skills |
| Strand 2. | Demonstrates understanding of movement concepts, principles, strategies, and tactics as they apply to the learning and performance of physical activities |
| | Concept 2: Scientific Principles |
| Strand 3. | Participates regularly in physical activity both during and beyond the structured physical education class |
| | Concept 2: Physical Activity Outside a Physical Education Program |
| Strand 4. | Achieves and maintains a health-enhancing level of physical fitness |
| | Concept 1: Health-Related Fitness |
| Strand 5. | Exhibits responsible personal and social behavior that respects self and others in physical activity settings |
| | Concept 1: Personal Behavior |
| | Concept 2: Social Behavior |
| Strand 6. | Values physical activity for health, enjoyment, challenge, self-expression, and/or social interaction |
| | Concept 1: Values Physical Activity |

**Table 6 – Arizona Grade 3 Visual Arts Concepts and Strands**

| Arizona Grade 3 Visual Arts Concepts and Strands | |
|---|---|
| Strand 1. | Create: Student will create artworks to communicate ideas, meanings, and/or purposes |
| | Concept 2: Materials, Tools, and Techniques |
| | Concept 3: Elements and Principles |
| Strand 2. | Relate: Student will analyze and interpret contextual ideas, meanings, and purposes of art from diverse cultures and time periods |
| | Concept 1: Art worlds |
| | Concept 2: Materials, Tools, and Techniques |
| | Concept 3: Elements and Principles |
| | Concept 4: Meanings or Purposes |
| | Concept 5: Quality |
| Strand 3. | Evaluate: Student will draw thoughtful conclusions about the significance of art |
| | Concept 2: Materials, Tools, and Techniques |
| | Concept 3: Elements and Principles |

**Table 7 – Arizona Grade 8 Visual Arts Concepts and Strands**

| Arizona Grade8 Visual Arts Concepts and Strands | |
|---|---|
| Strand 1. | Create: Student will create artworks to communicate ideas, meanings, and/or purposes<br>Concept 1: Creative Process<br>Concept 3: Elements and Principles<br>Concept 4: Meanings or Purposes<br>Concept 5: Quality |
| Strand 2. | Relate: Student will analyze and interpret contextual ideas, meanings, and purposes of art from diverse cultures and time periods<br>Concept 1: Art worlds<br>Concept 2: Materials, Tools, and Techniques<br>Concept 3: Elements and Principles<br>Concept 4: Meanings or Purposes<br>Concept 5: Quality |
| Strand 3. | Evaluate: Student will draw thoughtful conclusions about the significance of art<br>Concept 1: Arts Issues and Values<br>Concept 2: Materials, Tools, and Techniques<br>Concept 3: Elements and Principles<br>Concept 4: Meanings or Purposes<br>Concept 5: Quality |

**Table 8 – Arizona High School Visual Arts Concepts and Strands**

| Arizona High School Visual Arts Concepts and Strands | |
|---|---|
| Strand 1. | Create: Student will create artworks to communicate ideas, meanings, and/or purposes<br>Concept 2: Materials, Tools, and Techniques<br>Concept 3: Elements and Principles |
| Strand 2. | Relate: Student will analyze and interpret contextual ideas, meanings, and purposes of art from diverse cultures and time periods<br>Concept 1: Art worlds<br>Concept 2: Materials, Tools, and Techniques<br>Concept 3: Elements and Principles<br>Concept 4: Meanings or Purposes<br>Concept 5: Quality |
| Strand 3. | Evaluate: Student will draw thoughtful conclusions about the significance of art<br>Concept 1: Arts Issues and Values<br>Concept 2: Materials, Tools, and Techniques<br>Concept 3: Elements and Principles<br>Concept 4: Meanings or Purposes<br>Concept 5: Quality |

**Table 9 – Arizona Grade 3 Music Concepts and Strands**

| Arizona Grade 3 Music Concepts and Strands | |
|---|---|
| Strand 1. | Create |
| | Concept 5: Reading and notating music |
| Strand 2. | Relate |
| | Concept 1: Understanding the relationships among music, the arts & other disciplines outside the arts |
| | Concept 2: Understanding music in relation to history and culture |
| Strand 3. | Evaluate |
| | Concept 1: Listening to analyzing and describing music |

**Table 10 – Arizona Band Concepts and Strands**

| Arizona Grade 3 Music Concepts and Strands | |
|---|---|
| Strand 1. | Create |
| | Concept 2: Playing instruments alone and with others, music from different genres and diverse cultures |
| | Concept 5: Reading and notating music |
| Strand 2. | Relate |
| | Concept 1: Understanding the relationships among music, the arts & other disciplines outside the arts |
| | Concept 2: Understanding music in relation to history and culture |
| | Concept 3: Understanding music in relation to self and universal themes |
| Strand 3. | Evaluate |
| | Concept 1: Listening to analyzing and describing music |
| | Concept 2: Evaluating music and music performances |

**Table 11 – Arizona Choir Concepts and Strands**

| Arizona Choir Concepts and Strands | |
|---|---|
| Strand 1. | Create |
| | Concept 1: Singing alone & with others music from different genres and diverse cultures. |
| | Concept 4: Composing and arranging music |
| | Concept 5: Reading and notating music |
| Strand 2. | Relate |
| | Concept 1: Understanding the relationships among music, the arts & other disciplines outside the arts |
| | Concept 2: Understanding music in relation to history and culture |
| Strand 3. | Evaluate |
| | Concept 1: Listening to analyzing and describing music |
| | Concept 2: Evaluating music and music performances |

**Table 12 – Arizona Theatre Concepts and Strands**

| Arizona Theatre Concepts and Strands | |
|---|---|
| Strand 1. | Create - The processes and experiences developed related to theatre<br>Concept 1: Collaboration - Collaboration includes working jointly, cooperating, negotiating, and articulating ideas to reach consensus that forms theatrical art.<br>Concept 2: Acting - Acting is the process and art of representing a character in the classroom, on stage, or in other media.<br>Concept 3: Theatre Technology and Design - Theatre technology uses craft skills, knowledge of design, equipment, and materials to construct the elements necessary for the visual and aural aspects of production that serve the script and the action.<br>Concept 4: Playwriting - Playwriting is the process of conceptualizing, devising, improvising, developing, writing, and revising original written work for the stage and other media.<br>Concept 5: Directing - Directing is the process of conceptualizing, organizing, and leading a collaborative process with the intent of performance. |
| Strand 2. | Relate - How the human experience influences and is influenced by theatre<br>Concept 1: Collaboration - Collaboration includes working jointly, cooperating, negotiating, and articulating ideas to reach consensus that forms theatrical art.<br>Concept 2: Acting - Acting is the process and art of representing a character in the classroom, on stage, or in other media.<br>Concept 3: Theatre Technology and Design - Theatre technology uses craft skills, knowledge of design, equipment, and materials to construct the elements necessary for the visual and aural aspects of production that serve the script and the action.<br>Concept 4: Playwriting - Playwriting is the process of conceptualizing, devising, improvising, developing, writing, and revising original written work for the stage and other media.<br>Concept 5: Directing - Directing is the process of conceptualizing, organizing, and leading a collaborative process with the intent of performance. |
| Strand 3. | Evaluate - The informal and formal reflection and critical analysis to address and assess the qualities of theatre<br>Concept 2: Acting - Acting is the process and art of representing a character in the classroom, on stage, or in other media.<br>Concept 3: Theatre Technology and Design - Theatre technology uses craft skills, knowledge of design, equipment, and materials to construct the elements necessary for the visual and aural aspects of production that serve the script and the action.<br>Concept 4: Playwriting - Playwriting is the process of conceptualizing, devising, improvising, developing, writing, and revising original written work for the stage and other media. |

**Table 13 – Arizona Dance Concepts and Strands**

| Arizona Dance Concepts and Strands | |
|---|---|
| Strand 1. | Create - Students explore, demonstrate and apply the elements and techniques of dance<br>Concept 1: Body - Identify, demonstrate and analyze the use of the body for dance through an understanding of anatomy, kinesiology and basic movement principles.<br>Concept 2: Movement Skills – Identify, demonstrate and analyze basic movement skills in the exploration and performance of dance.<br>Concept 3: Elements of Dance - Identify, demonstrate and analyze the elements of dance.<br>Concept 4: Improvisation/Choreography - Identify, demonstrate, analyze and apply improvisational structures, choreographic processes, forms and principles.<br>Concept 5: Performance Values - Identify, demonstrate and analyze the aesthetic values inherent in dance.<br>Concept 6: Production/Design - Identify, demonstrate, analyze and apply the elements of theatrical production as they relate to dance. |
| Strand 2. | Relate – Students understand how the human experience influences and is influenced by dance, and apply dance to understand ideas across disciplines<br>Concept 1: Relating Dance Forms and History – Identify, demonstrate and analyze the origins, history and continuing evolution of various dance forms.<br>Concept 2: Relating Dance with Social and Cultural Influences - Identify, demonstrate and analyze the reciprocal relationships between dance and society. |
| Strand 3. | Evaluate - The informal and formal reflection and critical analysis to address and assess the qualities of theatre<br>Concept 1: Understanding Dance – Identify, reflect, analyze and interpret how dance communicates and conveys meaning. |

**Table 14 – Arizona Grade 6 Social Studies Concepts and Strands**

| Arizona Grade 6 Social Studies Concepts and Strands | |
|---|---|
| AZCCRS. | Reading<br>Key Ideas and Details<br>Craft and Structure<br>Integration of Knowledge and Ideas |
| Strand 1. | American History<br>Concept 2: Early Civilizations<br>Concept 10: Contemporary U.S. |
| Strand 2. | World History<br>Concept 1: Research Skills for History<br>Concept 2: Early Civilizations<br>Concept 3: World of Transition<br>Concept 4: Renaissance and Reformation<br>Concept 5: Encounters and Exchange |
| Strand 3. | Civics/Government<br>Concept 3: Functions of Government<br>Concept 5: Government Systems of the World |
| Strand 4. | Geography<br>Concept 1: World in Spatial Terms<br>Concept 2: Places and Regions<br>Concept 4: Human Systems<br>Concept 5: Environment and Society<br>Concept 6: Geographic Applications |
| Strand 5. | Economics<br>Concept 1: Foundations of Economics |

**Table 15 – Arizona Grade 7 Social Studies Concepts and Strands**

| Arizona Grade 7 Social Studies Concepts and Strands | |
|---|---|
| AZCCRS. | Reading |
| | Key Ideas and Details |
| | Craft and Structure |
| | Integration of Knowledge and Ideas |
| Strand 1. | American History |
| | Concept 1: Research Skills for History |
| | Concept 6: Civil War and Reconstruction |
| | Concept 7: Emergence of Modern U.S. |
| | Concept 8: Great Depression and WWII |
| Strand 2. | World History |
| | Concept 6: Age of Revolution |
| | Concept 7: Age of Imperialism |
| | Concept 8: World at War |
| Strand 3. | Civics/Government |
| | Concept 1: Foundations of Government |
| | Concept 2: Structure of Government |
| | Concept 3: Functions of Government |
| | Concept 4: Rights, Responsibilities, and Roles of Citizenship |
| | Concept 5: Government Systems of the World |
| Strand 4. | Geography |
| | Concept 1: World in Spatial Terms |
| | Concept 4: Human Systems |
| | Concept 5: Environment and Society |
| | Concept 6: Geographic Application |
| Strand 5. | Economics |
| | Concept 1: Foundations of Economics |
| | Concept 2: Microeconomics |
| | Concept 3: Macroeconomics |
| | Concept 4: Global Economics |
| | Concept 5: Personal Finance |

**Table 16 – Arizona Grade 8 Social Studies Concepts and Strands**

| Arizona Grade 8 Social Studies Concepts and Strands | |
|---|---|
| AZCCRS. | Reading |
| | Key Ideas and Details |
| | Craft and Structure |
| | Integration of Knowledge and Ideas |
| Strand 1. | American History |
| | Concept 1: Research Skills for History |
| | Concept 4: Revolution and a New Nation |
| | Concept 8: Great Depression and WWII |
| | Concept 9: Postwar U.S. |
| | Concept 10: Contemporary U.S. |
| Strand 2. | World History |
| | Concept 1: Research Skills for History |
| | Concept 8: World at War |
| Strand 3. | Civics/Government |
| | Concept 1: Foundations of Government |
| | Concept 2: Structure of Government |
| | Concept 3: Functions of Government |
| | Concept 4: Rights, Responsibilities, and Roles of Citizenship |
| | Concept 5: Government Systems of the World |
| Strand 4. | Geography |
| | Concept 1: World in Spatial Terms |
| | Concept 2: Places and Regions |
| | Concept 4: Human Systems |
| Strand 5. | Economics |
| | Concept 1: Foundations of Economics |
| | Concept 4: Global Economics |
| | Concept 5: Personal Finance |

**Table 17 – Arizona U.S. History Concepts and Strands**

| Arizona U.S. History Concepts and Strands | |
|---|---|
| AZCCRS. | Reading |
| | Key Ideas and Details |
| | Craft and Structure |
| Strand 1. | American History |
| | Concept 1: Research Skills for History |
| | Concept 3: Exploration and Colonization |
| | Concept 4: Revolution and a New Nation |
| | Concept 5: Westward Expansion |
| | Concept 6: Civil War and Reconstruction |
| | Concept 7: Emergence of the Modern U.S. |
| | Concept 8: Great Depression and WWII |
| | Concept 9: Postwar U.S. |
| Strand 4. | Geography |
| | Concept 5: Environment and Society |

**Table 18 – Arizona World History Concepts and Strands**

| Arizona World History Concepts and Strands | |
|---|---|
| AZCCRS. | Reading |
| | Key Ideas and Details |
| | Craft and Structure |
| Strand 2. | World History |
| | Concept 1: Research Skills for History |
| | Concept 2: Early Civilizations |
| | Concept 3: World of Transition |
| | Concept 4: Renaissance and Reformation |
| | Concept 5: Encounters and Exchange |
| | Concept 6: Age of Revolution |
| | Concept 7: Age of Imperialism |
| | Concept 8: World at War |
| | Concept 9: Contemporary World |
| Strand 4. | Geography |
| | Concept 4: Human Systems |

**Table 19 – Arizona Economics Concepts and Strands**

| Arizona Economics Concepts and Strands | |
|---|---|
| AZCCRS. | Reading |
| | Key Ideas and Details |
| | Craft and Structure |
| Strand 4. | Geography |
| | Concept 2: Places and Regions |
| Strand 5. | Economics |
| | Concept 1: Foundations of Economics |
| | Concept 2: Microeconomics |
| | Concept 3: Macroeconomics |
| | Concept 4: Global Economics |
| | Concept 5: Personal Finance |

**Table 20 – Arizona Civics/Government Concepts and Strands**

| Arizona Civics/Government Concepts and Strands | |
|---|---|
| AZCCRS. | Reading |
| | Key Ideas and Details |
| | Craft and Structure |
| Strand 3. | Civics/Government |
| | Concept 1: Foundations of Government |
| | Concept 2: Structure of Government |
| | Concept 3: Functions of Government |
| | Concept 4: Rights, Responsibilities, and Roles of Citizenship |
| | Concept 5: Government Systems of the World |
| Strand 4. | Geography |
| | Concept 1: World in Spatial Terms |
| | Concept 2: Places and Regions |

## 3.2 Test Blueprints

A test blueprint designates the percentage of items that should measure each strand and concept. The pre- and post-assessments were designed in accordance with the blueprints shown in Tables 21 through 34. The test blueprints were used with the processes described in detail in Part 4 of this technical manual to develop all field tests and operational tests.

**Table 21 – MCESA Blueprint for Physical Education Grades 3, 8, and 9-12**

| Reporting Category | Grade 3 | Grade 8 | Grades 9–12 |
|---|---|---|---|
| Strand 1 | 10% | 16% | 20% |
| Strand 2 | 40% | 31% | 35% |
| Strand 3 | 0% | 8% | 7% |
| Strand 4 | 30% | 16% | 11% |
| Strand 5 | 20% | 18% | 7% |
| Strand 6 | 0% | 11% | 20% |
| | 100% | 100% | 100% |

**Table 22 – Blueprint by Strand and Concept Level for Physical Education Grades 3, 8, and 9-12**

| | Grade Level | | |
|---|---|---|---|
| **Strand 1** | **3** | **8** | **9–12** |
| Concept 1: Fundamental Movement Skills | 10% | 9% | 0% |
| Concept 2: Rhythmic Movement | 0% | 0% | 0% |
| Concept 3: Complex or Specialized Movement Skills | 0% | 7% | 20% |
| **Strand 2** | | | |
| Concept 1: Movement Concepts | 9% | 9% | 0% |
| Concept 2: Scientific Principles | 22% | 13% | 35% |
| Concept 3: Strategies and Tactics | 9% | 9% | 0% |
| **Strand 3** | | | |
| Concept 1: Physical Activity in PE Program | 0% | 0% | 0% |
| Concept 2: Physical Activity Outside PE Program | 0% | 8% | 7% |
| **Strand 4** | | | |
| Concept 1: Health-Related Fitness | 30% | 16% | 11% |
| **Strand 5** | | | |
| Concept 1: Personal Behavior | 9% | 7% | 4% |
| Concept 2: Social Behavior | 11% | 11% | 3% |
| **Strand 6** | | | |
| Concept 1: Values Physical Activity | 0% | 11% | 20% |

**Table 23 – MCESA Blueprint for Visual Arts Grades 3, 8, and 9-12**

| Reporting Category | Grade 3 | Grade 8 | Grades 9–12 |
|---|---|---|---|
| Strand 1: Create | 18% | 22% | 15% |
| Strand 2: Relate | 53% | 40% | 35% |
| Strand 3: Evaluate | 29% | 38% | 50% |
| | 100% | 100% | 100% |

**Table 24 - Blueprint by Strand and Concept Level for Visuals Arts Grades 3, 8, and 9-12**

| | Grade Level | | |
|---|---|---|---|
| **Strand 1: Create** | **3** | **8** | **12** |
| Concept 1: Creative Process | 0% | 3% | 0% |
| Concept 2: Materials, Tools, & Techniques | 9% | 0% | 6% |
| Concept 3: Elements and Principles | 9% | 11% | 9% |
| Concept 4: Meanings or Purposes | 0% | 4% | 0% |
| Concept 5: Quality | 0% | 4% | 0% |
| **Strand 2: Relate** | | | |
| Concept 1: Art worlds | 9% | 18% | 4% |
| Concept 2: Materials, Tools, and Techniques | 17% | 11% | 8% |
| Concept 3: Elements and Principles | 9% | 2% | 4% |
| Concept 4: Meanings or Purposes | 9% | 7% | 17% |
| Concept 5: Quality | 9% | 2% | 2% |
| **Strand 3: Evaluate** | | | |
| Concept 1: Arts Issues and Values | 0% | 4% | 4% |
| Concept 2: Materials, Tools, and Techniques | 9% | 4% | 11% |
| Concept 3: Elements and Principles | 20% | 21% | 31% |
| Concept 4: Meanings or Purposes | 0% | 2% | 2% |
| Concept 5: Quality | 0% | 7% | 2% |

**Table 25 – MCESA Blueprint for Grade 3 Music, Choir, and Band**

| Reporting Category | Grade 3 | Choir | Band |
|---|---|---|---|
| Strand 1: Create | 56% | 42% | 67% |
| Strand 2: Relate | 22% | 29% | 20% |
| Strand 3: Evaluate | 22% | 29% | 13% |
| | 100% | 100% | 100% |

**Table 26 - MCESA Blueprint by Strand and Concept for Grade 3 Music, Choir, and Band**

| | Course | | |
|---|---|---|---|
| **Strand 1: Create** | **Grade 3** | **Choir** | **Band** |
| Concept 1: Singing, alone and with others . . . | 0% | 9% | 0% |
| Concept 2: Playing instruments . . . | 0% | 0% | 27% |
| Concept 3: Improvising rhythms, melodies, variations, and accompaniments | 0% | 0% | 0% |
| Concept 4: Composing and arranging music | 0% | 2% | 0% |
| Concept 5: Reading and notating music | 56% | 31% | 40% |
| **Strand 2: Relate** | **Grade 3** | **Choir** | **Band** |
| Concept 1: Understanding the relationships among music, the arts, and other disciplines | 11% | 18% | 14% |
| Concept 2: Understanding music in relation to history and culture | 11% | 11% | 4% |
| Concept 3: Understanding music in relation to self and universal themes | 0% | 0% | 2% |
| **Strand 3: Evaluate** | **Grade 3** | **Choir** | **Band** |
| Concept 1: Listening to, analyzing, describing music | 22% | 20% | 11% |
| Concept 2: Evaluating music and music performances | 0% | 9% | 2% |

**Table 27 – MCESA Blueprint for High School Theater**

| **Reporting Category** | **Grades 9–12** |
|---|---|
| Strand 1: Create | 56% |
| Strand 2: Relate | 33% |
| Strand 3: Evaluate | 11% |
| | 100% |

**Table 28 – MCESA Blueprint by Strand and Concept for High School Theater**

| **Strand 1: Create** | **Grades 9–12** |
|---|---|
| Concept 1: Collaboration | 11% |
| Concept 2: Acting | 16% |
| Concept 3: Technical Theater and Design | 16% |
| Concept 4: Playwriting | 11% |
| Concept 5: Directing | 2% |
| **Strand 2: Relate** | |
| Concept 1: Collaboration | 4% |
| Concept 2: Acting | 16% |
| Concept 3: Technical Theater and Design | 2% |
| Concept 4: Playwriting | 7% |
| Concept 5: Directing | 4% |
| **Strand 3: Evaluate** | |
| Concept 1: Collaboration | 0% |
| Concept 2: Acting | 5% |
| Concept 3: Technical Theater and Design | 2% |
| Concept 4: Playwriting | 4% |
| Concept 5: Directing | 0% |

**Table 29 – MCESA Blueprint for High School Dance**

| Reporting Category | Grades 9–12 |
|---|---|
| Strand 1: Create | 80% |
| Strand 2: Relate | 12% |
| Strand 3: Evaluate | 8% |
| | 100% |

**Table 30 – MCESA Blueprint by Strand and Concept for High School Dance**

| Strand 1: Create | Grades 9–12 |
|---|---|
| Concept 1: Body | 23% |
| Concept 2: Movement Skills | 23% |
| Concept 3: Elements of Dance | 13% |
| Concept 4: Improvisation/Choreography | 13% |
| Concept 5: Performance Values | 4% |
| Concept 6: Production/Design | 4% |
| **Strand 2: Relate** | **Grades 9–12** |
| Concept 1: Relating Dance Forms and History | 11% |
| Concept 2: Relating Dance with Social and Cultural Influences | 2% |
| Concept 3: Relating Dance and Literacy | 0% |
| Concept 4: Relating Dance with other Disciplines | 0% |
| Concept 5: Relating Dance and Music | 0% |
| **Strand 3: Evaluate** | **Grades 9–12** |
| Concept 1: Understanding Dance | 8% |
| Concept 2: Professionalism | 0% |

**Table 31 – MCESA Blueprint for Social Studies Grades 6, 7, 8**

| Reporting Category | Grade 6 | Grade 7 | Grade 8 |
|---|---|---|---|
| AZCC Strand: Reading | 10% | 10% | 10% |
| AZCC Strand: Writing | 0% | 0% | 0% |
| Strand 1: American History | 10% | 40% | 30% |
| Strand 2: World History | 40% | 10% | 20% |
| Strand 3: Civics/Government | 10% | 20% | 20% |
| Strand 4: Geography | 20% | 10% | 10% |
| Strand 5: Economics | 10% | 10% | 10% |
| | 100% | 100% | 100% |

**Table 32 – MCESA Blueprint by Strand and Concept for Social Studies Grades 6, 7, 8**

| **AZCC Strand: Reading** | **Grade 6** | **Grade 7** | **Grade 8** |
|---|---|---|---|
| Key Ideas and Details | 4% | 4% | 4% |
| Craft and Structure | 4% | 4% | 4% |
| Integration of Knowledge and Ideas | 2% | 2% | 2% |
| Range of Reading and Text Complexity | 0% | 0% | 0% |
| **Strand 1: American History** | **Grade 6** | **Grade 7** | **Grade 8** |
| Concept 1: Research Skills for History | 0% | 10% | 5% |
| Concept 2: Early Civilizations | 7% | 0% | 0% |
| Concept 3: Exploration and Colonization | 0% | 0% | 0% |
| Concept 4: Revolution and a New Nation | 0% | 0% | 7% |
| Concept 5: Westward Expansion | 0% | 0% | 0% |
| Concept 6: Civil War and Reconstruction | 0% | 7% | 0% |
| Concept 7: Emergence of the Modern US | 0% | 18% | 0% |
| Concept 8: Great Depression and WWII | 0% | 5% | 6% |
| Concept 9: Postwar U.S. | 0% | 0% | 7% |
| Concept 10: Contemporary U.S. | 3% | 0% | 5% |
| **Strand 2: World History** | **Grade 6** | **Grade 7** | **Grade 8** |
| Concept 1: Research Skills for History | 10% | 0% | 5% |
| Concept 2: Early Civilizations | 17% | 0% | 0% |
| Concept 3: World In Transition | 8% | 0% | 0% |
| Concept 4: Renaissance and Reformation | 3% | 0% | 0% |
| Concept 5: Encounters and Exchange | 2% | 0% | 0% |
| Concept 6: Age of Revolution | 0% | 2% | 0% |
| Concept 7: Age of Imperialism | 0% | 5% | 0% |
| Concept 8: World at War | 0% | 3% | 15% |
| Concept 9: Contemporary World | 0% | 0% | 0% |
| **Strand 3: Civics/Government** | **Grade 6** | **Grade 7** | **Grade 8** |
| Concept 1: Foundations of Government | 0% | 3% | 3% |
| Concept 2: Structure of Government | 0% | 4% | 5% |
| Concept 3: Functions of Government | 7% | 7% | 7% |
| Concept 4: Rights, Responsibilities, Roles of Citizenship | 0% | 3% | 2% |
| Concept 5: Government Systems of the World | 3% | 3% | 3% |
| **Strand 4: Geography** | **Grade 6** | **Grade 7** | **Grade 8** |
| Concept 1: The World in Spatial Terms | 7% | 3% | 2% |
| Concept 2: Places and Regions | 2% | 0% | 5% |
| Concept 3: Physical Systems | 0% | 0% | 0% |
| Concept 4: Human Systems | 5% | 3% | 3% |
| Concept 5: Environment and Society | 5% | 2% | 0% |
| Concept 6: Geographic Applications | 1% | 2% | 0% |
| **Strand 5: Economics** | **Grade 6** | **Grade 7** | **Grade 8** |
| Concept 1: Foundations of Economics | 10% | 2% | 4% |
| Concept 2: Microeconomics | 0% | 3% | 0% |
| Concept 3: Macroeconomics | 0% | 2% | 0% |
| Concept 4: Global Economics | 0% | 1% | 3% |
| Concept 5: Personal Finance | 0% | 2% | 3% |

**Table 33 – MCESA Blueprint for U.S. History, World History, Civics/Government, and Economics**

| Reporting Category | U.S. History | World History | Civics/Gov't | Economics |
|---|---|---|---|---|
| CCSS Strand: Reading | 10% | 10% | 10% | 10% |
| CCSS Strand: Writing | 0% | 0% | 0% | 0% |
| Strand 1: American History | 80% | 0% | 0% | 0% |
| Strand 2: World History | 0% | 80% | 0% | 0% |
| Strand 3: Civics/Government | 0% | 0% | 80% | 0% |
| Strand 4: Geography | 10% | 10% | 10% | 10% |
| Strand 5: Economics | 0% | 0% | 0% | 80% |
| Total | 100% | 100% | 100% | 100% |

**Table 34 – MCESA Blueprint by Strand and Concept for U.S. History, World History, Civics/Government, and Economics**

| U.S. History | % of Items |
|---|---|
| Common Core State Standards for ELA and Literacy in History/Social Studies | 10% |
| Strand 1 Concept 1: Research | 10% |
| Strand 1 Concept 2: Early Civilizations | 0% |
| Strand 1 Concept 3: Exploration and Colonization | 10% |
| Strand 1 Concept 4: Revolution and a New Nation | 10% |
| Strand 1 Concept 5: Westward Expansion | 10% |
| Strand 1 Concept 6: Civil War and Reconstruction | 10% |
| Strand 1 Concept 7: Emergence of the Modern U.S. | 10% |
| Strand 1 Concept 8: Great Depression and WWII | 10% |
| Strand 1 Concept 9: Postwar U.S. | 10% |
| Strand 1 Concept 10: Contemporary U.S. | 0% |
| Strand 4 Concept 5: Environment and Society | 10% |
| **World History** | **% of Items** |
| Common Core State Standards for ELA and Literacy in History/Social Studies | 10% |
| Strand 2 Concept 1: Research Skills for History | 10% |
| Strand 2 Concept 2: Early Civilizations | 10% |
| Strand 2 Concept 3: World in Transition | 10% |
| Strand 2 Concept 4: Renaissance and Reformation<br>Strand 2 Concept 5: Encounters and Exchange | 15% |
| Strand 2 Concept 6: Age of Revolution<br>Strand 2 Concept 7: Age of Imperialism | 15% |
| Strand 2 Concept 8: World at War | 10% |
| Strand 2 Concept 9: Contemporary World | 10% |
| Strand 4 Concept 4: Human Systems | 10% |

| **Civics/Government** | **% of Items** |
|---|---|
| Common Core State Standards for ELA and Literacy in History/Social Studies | 10% |
| Strand 3 Concept 1: Foundations of Government | 15% |
| Strand 3 Concept 2: Structure of Government | 15% |
| Strand 3 Concept 3: Functions of Government | 15% |
| Strand 3 Concept 4: Rights, Responsibilities, and Roles of Citizenship | 25% |
| Strand 3 Concept 5: Government Systems of the World | 10% |
| Strand 4 Concept 1: World in Spatial Terms<br>Strand 4 Concept 2: Places and Regions | 10% |
| **Economics** | **% of Items** |
| Common Core State Standards for ELA and Literacy in History/Social Studies | 10% |
| Strand 5 Concept 1: Foundation of Economics | 15% |
| Strand 5 Concept 2: Microeconomics | 20% |
| Strand 5 Concept 3: Macroeconomics | 20% |
| Strand 5 Concept 4: Global Economics | 10% |
| Strand 5 Concept 5: Personal Finance | 15% |
| Strand 4 Concept 2: Places and Regions | 10% |

### 3.3 Test Design

The original test design called for the development of enough items to create two operational summative test forms. The forms would include multiple-choice items that measure end-of-course content. In order to create two operational summative test forms, three field-test forms, each with 45 items, would need to be administered. This number of items per form was determined by the minimum length of class time that students spend in a class period and by the fact that multiple-choice items take, on average, about one minute per item.

The final number of field-test forms (three forms: A, B, and C) was determined by the sample size of students estimated to participate in the field testing. The plan was to field test approximately 1,500 student responses for each item in order to generate three-parameter Item Response Theory (IRT) data. The number of forms was also determined by the number of field-test items needed to create two operational forms and include overage in the event that items were not deemed usable on the operational forms because they did not meet item parameters.

The Spring, 2012 Phase I summative field tests followed this design of three forms with 45 items per form. However, with the addition of the pre-assessment requirements needed to show growth, the number of items per form was increased to 50 for the Phase II and Phase III content areas. This increase was due to the addition of anchor items that linked the pre- and post-assessments along with the linking items between forms, which reduced the number of items being field tested. The final test design resulted in all operational test forms for phases I, II and III having 50 items. The social studies assessment included a reading passage with five of the items aligned to the passage.

The pre-assessment forms were designed to be summative, but not entirely comprised of end-of-course content. The pre-assessments have a 40-60 design, whereby 60% of the test content is

aligned to end-of-course content, while 40% is aligned to entry-level course content. This design is intended to increase the validity of the pre-assessment to align to the students' abilities and the course constructs at the beginning of the course. The items used for the 40% portion of the pre-assessment align to item specifications written for this purpose.

Pre-assessment field tests were administered at the beginning of each course, during the same window in which the operational pre-assessments would be administered, in order to ensure that the field-test item statistics would be valid to use for the operational pre-assessments. Similarly, post-assessment field tests were administered at the end of each course, during the same window in which the operational post-assessment tests would be administered, in order to ensure that the field-test item statistics would be valid to use for the operational post-assessments.

# Part 4. Test Development

## 4.1 Test Development and Editing Processes

Several different processes were employed throughout the development of the MCESA Content Specific Assessments. Some process involved committees of educators while others were done internally by WestEd or MCESA.

### *4.1.1 General Test Development Process*

Test development for the MCESA Content Specific Assessments began with a project kick-off meeting in January 2012. The project deliverables were defined, including online tests, answer documents, and test administration manuals. The process to produce the online field and operational pre- and post-assessments was also defined. Pre- and post-assessment blueprints and item specifications were developed at committee meetings. These blueprints and item specifications, along with the style guide for Arizona's Instrument to Measure Standards (AIMS) (customized to MCESA style for online testing), were used by item writers to produce the entry and summative items. Items were then reviewed by WestEd and MCESA editors, the bias/sensitivity review committees, and item content review committees. Accepted entry and summative items were field tested in three field-test forms per pre- or post-assessment, from which operational forms were produced.

### *4.1.2 Blueprint and Item Specification Process*

Prior to item writing, county educators, with the aid of WestEd or MCESA facilitators, wrote blueprint and item specifications documents. The blueprints narrowed the content of the course to be assessed and defined the construct to be assessed. The item specifications provided definitions of what is tested by each PO, clarification of the PO statements, the content limits, the stimulus and response attribute descriptions, and the cognitive levels (DOK) of the POs. They help to inform item writing by explaining, in detail, what each PO means at each grade and subject and by describing how each PO is to be tested.

The blueprint and item specifications development committee used the following process:

- The committee determined whether each performance objective should be considered multiple choice, performance based, or classroom only.
- The committee determined the percentages of emphasis for the multiple-choice items across the reporting categories.
- The committee determined the number of items that should be tested for each concept for each form.
- The committee cross-checked percentages between grades of similar subjects, including:
    - checking for logical progressions from grade to grade, and
    - checking for proper emphasis from grade to grade.

- The committee developed item specifications for the selected POs being assessed as multiple-choice. This development included the following:
    - identifying limitations ("i.e.") in the POs;
    - identifying examples ("e.g.") in the POs;
    - defining terms in the POs; and
    - listing materials/tools/equipment for the POs.
- The committee assigned a cognitive level (DOK) to each PO.
    - WestEd staff trained the committee in determining the DOK level (1, 2, or 3) of POs.

WestEd and/or MCESA reviewed the committees' completed item specifications and made additional changes and corrections where needed. It was emphasized that the item specifications were living documents that could be updated at any time throughout the development process.

### *4.1.3 Item Writing Process*

The item writing process involved collaboration among many professionals and educators from WestEd and MCESA, in an effort to ensure that all developed items closely matched the Arizona content standards and the item specifications, and that enough items were developed to produce two comparable operational forms. The educators who were selected to serve on the item writing committee all possessed content and assessment expertise and the ability to be creative while adhering to the test blueprints, detailed item specifications, and content limitations.

Professional development was provided to integrate new participants with more experienced item writers. The educators wrote test items using templates that contained all item requirements and supporting information, such as strand, concept, performance objective, and content reference documentation.

Facilitators were assigned to each content area during the item writing workshops. The facilitators provided regular feedback to item writers, and discussed common issues that arose during the item-writing process with them. All item writers were required to sign a statement in which they agreed to treat all materials and communication related to item development as confidential.

The item writing committee activities included the following:

- The item writers were trained on how to write multiple-choice items, including instruction on:
    - defining the parts of a multiple-choice item (stem, correct response, distracters, and graphics);
    - how to locate and document graphics and passages;
    - how to use graphic templates;
    - the need to write depth of knowledge (DOK) level 2 items;
    - correct use of the MCESA style guide;

  - how to enter items into electronic content management systems; and
  - preferred best practices in item writing.
    - The multiple-choice items had four options, with one and only one correct answer.
    - Items do not use "all of the above," "none of the above," or answer options similar to these constructions.
    - Answer response options were parallel in structure and length.

- The item writers received a detailed item order. The detailed item order specified the number and types of items to be developed to the item specifications for that grade and subject, and any targeted cognitive complexity and difficulty.
- The item writers wrote items, using the item specifications for each grade and subject.
  - The item writers set up systems, within their grade/subject subcommittees, to keep track of item topics and to avoid duplication of items in the same POs.
  - The item writers also set up systems to keep track of names used, to avoid duplication and to avoid bias issues.
  - The item writers assigned a cognitive level (DOK level 1, 2, or 3) to each item.
  - The item writers assigned an item difficulty to each item (easy, medium, or hard).
- For items requiring graphics, the item writer submitted a graphic request to the facilitator.
- The theater educators selected and reviewed excerpts from plays for the high school theater assessment.
- After an item writer submitted an item, the facilitator provided a quick review to determine whether the item was viable or needed improvement. Quick feedback to writers provided important, timely information to build and sharpen their skills, particularly with respect to alignment and adherence to the item specifications. When an item was returned to the writer, the item writer reviewed the notes from the facilitator, completed the revision, and resubmitted the item.

For Phases I and II, a total of 168 items per assessment were developed. For Phase III, 150 items were developed. These totals represented the required number of items needed to create three field-test forms including overage to accommodate items that would be eliminated in the review and field-test processes.

### *4.1.4 Theater Excerpt Reviews*

The theater educators reviewed excerpts that had been preselected by WestEd content staff, including content and bias reviews, and approved or rejected them. The theater educators selected additional excerpts and reviewed them for content and bias.

A WestEd content specialist conducted the initial search for theater excerpts. The resulting set of excerpts was then presented to the theater item writing committee for consideration for use. The

committee accepted or rejected the excerpts and then wrote items to the accepted excerpts. Since the committee also searched for excerpts, committee members were trained using the same information and guidelines that the WestEd content specialist used to select excerpts.

The following information and guidelines were used to select excerpts:

- a list of the 12 theater POs that indicated the use of excerpts to test the PO;
- the theater item specifications;
- guidelines on unacceptable and avoidable topics; and
- other guidelines used by those selecting or writing excerpts.

Topics were deemed unacceptable for any of the following reasons:

- The topic could evoke unpleasant emotions or create humor for test takers, which might hamper their ability to take the remainder of the test in the optimal frame of mind.
- The topic is controversial among the adult population and might not be acceptable in a testing situation.
- The topic has been used extensively in standardized tests or textbooks, which might make it overly familiar and/or boring to students.
- The topic could appear biased against or toward some group of people.

The following list of avoidable topics was intended to address the primary topic of an excerpt (for example, an excerpt might be set during the time of the Revolutionary War, but it may be acceptable if the emphasis of the excerpt is not on the war itself):

- alcoholic beverages, tobacco, or drugs
- birthdays
- bodily functions
- celebrities still living (including athletes, actors, politicians, musicians, etc.)
- catastrophes/disasters
- children dealing with a serious issue
- computers in the home (use of computers is an acceptable topic in a school or public-library setting)
- creatures from outer space
- dancing that could be considered inappropriate
- death
- dinosaurs and prehistoric times; geological history (dependent on state guidelines)

- diseases
- evolution
- expensive gifts, vacations, or prizes
- gambling
- Halloween
- homes with swimming pools
- junk food (acceptable if in regard to healthier diet/lifestyle choices)
- movies
- nuclear weapons
- parapsychology
- politics
- religion, including religious holidays
- rock-and-roll music
- sex
- slavery
- violence
- war and bloodshed
- weapons (guns, knives, etc.)
- witchcraft, sorcery, etc.

Additionally, the committee was instructed to avoid any topic that may be:

- anthropomorphic (involving attribution of human characteristics to inanimate objects, animals, or natural phenomena; usually acceptable for classic folktale/fable/myth genres);
- critical of democracy or capitalism;
- dangerous for children (e.g., alone at home; swimming without adult supervision);
- demeaning to any group;
- disrespectful to authority or authority figures;
- highly controversial;
- involving middle-class amenities that may be unfamiliar to some children;

- regionalistic;
- smug, moralistic, or preachy; or
- stridently feminist or chauvinist.

Other guidelines used by those selecting or writing excerpts included:

- Choose or write excerpts that are short enough in length to show the excerpt and attached item(s) on a computer screen.
- Do not consider excerpts from plays that are typically used in classroom curricula

In all, 26 theater excerpts were selected as available to be used to assess 12 theater POs. Out of the 26 excerpts selected or written, nine were used in the theater pre-assessment field tests, and 11 were used in the theater post-assessment field tests.

### *4.1.5 Social Studies Passage Reviews*

At the blueprint development stage, MCESA decided to include items aligned to the Arizona College and Career Ready Standards for Literacy in History and Social Studies to the Phase III social studies pre- and post-assessments. Passages were selected by WestEd content specialists and reviewed and accepted or rejected by MCESA social studies and reading experts, using criteria such as reading level, alignment to Arizona standards, and bias/sensitivity.

The following information and guidelines were used to select the passages:

- a list of the Arizona Common Core for Reading POs for literacy in history and social studies that indicated where passages and items would be needed;
- the Phase III item specifications;
- the definition of informational text as defined by the Common Core State Standards range of test types, including subgenres of exposition, argument, and functional text in the forms of:
    - personal essays;
    - speeches;
    - opinion pieces;
    - essays about art or literature;
    - biographies;
    - memoirs;
    - journalism; and
    - historical, scientific, technical, or economic accounts;
- Common Core for Reading passage specifications (shown in Table 35).

**Table 35 – Common Core for Reading Passage Specifications**

| Grade | CCSS Lexile Range | Word Count Range |
|---|---|---|
| 6 | 955–1155 | 400-800 |
| 7 | 955–1155 | 500-950 |
| 8 | 955–1155 | 500-1000 |
| 9 | 1080–1305 | 500-1200 |
| 10 | 1080–1305 | 500-1200 |
| 11 | 1215–1355 | 500-1200 |

WestEd selected, and MCESA approved, 40 pre-assessment passages and 44 post-assessment passages across all of the social studies grades and subjects.

### *4.1.6 Item Editing and Graphic Development Processes*

Item editors shaped the items produced by item writers into more polished products. Their content expertise and assessment knowledge ensured that the items conformed to the rigorous content and style guidelines required. Several rounds of editing, identified as intake, E1, E2, proofreading and final-eye review, were applied, consistent with advancing levels of proficiency of the editorial staff.

A well-trained team of a coordinator, content leads, desktop publishing professionals, and proofreaders supported the editors. The coordinator maintained the project calendar and monitored the completion of writing, editing, and proofreading assignments. In this monitoring role, the coordinator was in close communication with the content leads regarding the flow of items throughout the writing and editorial process. The desktop publishing professionals created the graphics required for items, following exacting specifications to ensure content integrity and adherence to the specifications outlined in the style guide. The desktop publishers and proofreaders received training on MCESA style guide so that items delivered for review reflected the expected style and accuracy.

Desktop publishers began to interact with particular items after being notified by the coordinator that items had passed through intake. The desktop publishers followed the writers' instructions for creating the graphics, and contacted the coordinator or content lead to obtain clarification or discuss a detail of the graphic if needed. WestEd's desktop publishers possessed a strong sense of grade-level appropriateness and were encouraged to raise clarifying questions with content staff to ensure the optimal rendering of graphics that are true to the assessment anchors. WestEd's desktop publishers used the Adobe Creative Suite™, which included the most current versions of Adobe Illustrator, Adobe InDesign, and Adobe Photoshop, as well as Adobe Acrobat Professional Version.

Each item was edited with its associated metadata, including a unique item number and the content area, grade level, standard, concept, performance objective, item type, cognitive complexity, estimated difficulty, and answer key. The content experts developing the items were trained experts who ensured that items were grade-level-appropriate and consistent with Universal Design.

WestEd's development processes were informed by, and its practices refined to reflect, the elements of Universal Design that characterize sound assessment practice. The principles of Universal Design were created to ensure accessible environments for all people through equitable use, simple and intuitive design, effective communication, tolerance for variability, and minimal fatigue. Their application is defended by research that links them to higher performance for all students.

The National Center for Educational Outcomes has published guidelines for Universal Design (Thompson, et al., 2002b). WestEd incorporated these principles in both the development of items and the layout of test forms. Editors and desktop publishers were trained in applying Universal Design principles. Universal Design practices for item writing and editing included the following:

- using consistent naming and graphics conventions;
- replacing low-frequency words with simple common words;
- avoiding irregularly spelled words, words with ambiguous or multiple meanings, technical terms (unless defined and integral to meaning), and concepts with multiple names, symbols, or representations;
- ensuring clarity of noun-pronoun relationships; and
- simplifying keys and legends.

The complete editing and graphic development process included these steps.

- Intake
- Item editing (E1)
- Item editing (E2)
- Proofreading
- Final-eye review

The intake editor reviewed all of the fields that an item writer was required to complete and determined whether the item met the standard of quality. For each item, the intake editor:

- reviewed the item for alignment;
- checked that the necessary graphics and sources for data used within the item were provided;
- reviewed the item to ensure that it was within the parameters set by the style guide and the writer guidelines;
- reviewed the item for rigor, grade, and language appropriateness;
- reviewed item graphics if necessary; and
- evaluated the item for bias, sensitivity, and Universal Design issues.

The intake editor also provided the coordinator with general feedback. At the same time, the coordinator:

- worked with the intake editor to determine the intake priority, if needed (e.g., graphics items needed before items without graphics);
- ensured that any received intake graphics were sent to the intake editor, including any revised graphics; and
- delivered the graphics requests to desktop publishing staff.

After the intake review, accepted items moved to the E1 level for further editing. An intake note was written to the E1 editor with any suggestions for edits.

The E1 editor completed the first thorough review and edit of the item, ensuring that the item was in the correct format and met the expectations noted within the item guidelines and the style guide. The E1 editor performed the following tasks:

- reviewed the notes from the intake editor;
- reviewed the alignment of the item;
- checked the item to ensure that it was within the parameters set by the style guide and the item guidelines;
- checked the item for rigor, grade, and language appropriateness;
- formatted the item according to MCESA specifications for font, spacing, etc.;
- checked the answer choice and distracter rationales, as appropriate;
- reviewed the stem wording;
- checked the graphics for completeness and accuracy;
- verified references, as needed; and
- made a note about the item for the next editor, as needed.

The item editors used EDL Core Vocabularies in Reading, Mathematics, Science, and Social Studies and/or the Children's Writer's Word Book to ensure the use of grade-level-appropriate words in the items. In passages where words were at a higher grade level, the off-grade-level words were glossed and defined in footnotes

If needed, the E1 editor requested a graphics revision from desktop publishing and flagged the item appropriately. Once the graphic was revised, the item was returned to the editor and the editor reviewed the graphic before sending the item to the next level.

After an item had gone through E1 editing, the E1 editor sent the item to the E2 level within the item management system. The E2 editor examined the item closely, focusing effort on the details of the item's language. In addition, the E2 editor performed the following tasks:

- reviewed the notes from the intake editor and the E1 editor;

- reviewed the assigned standard, cognitive complexity level, and difficulty for
- alignment with item specifications;
- checked the item's match to the parameters established in the style guide and the item guidelines;
- checked the item for rigor, grade, and language appropriateness;
- checked the answer choices and the balance of answers, ensuring that there was only one correct answer;
- evaluated the wording of the item for use of clear, precise, and concise language;
- checked the graphics for completeness and accuracy;
- evaluated the item for bias, sensitivity, and fairness issues;
- evaluated the item for adherence to Universal Design principles; and
- made a note about the item for the next editor, as needed.

If needed, the E2 editor requested a graphics revision from desktop publishing, using the process followed by the E1 editor.

Once E2 edits were complete, the editor sent the item to the Proofreading status within the item management system. The purpose of proofreading items is to check for any errors such as spelling or grammatical mistakes. The coordinator, content lead, and proofreader were the key staff for this step of item development. The proofreader ensured that any errors were identified, checking:

- spelling and grammar;
- item card formatting for paper review;
- match between the rationales and the correct answer key;
- for style errors;
- for content errors; and
- size, scale, and format of graphics.

After an item was proofread, the proofreader initialed the item page and returned the item to the coordinator.

The final-eye review stage was the last review of an item before it was tagged as ready for delivery for the bias and item content review meetings. The purpose of the final-eye stage was to ensure that each item followed the specified style and was correct and sound with respect to content. Items that did not pass the final-eye stage were edited as needed, as an edit level three (E3), and then received final-eye review and sign-off.

At the final-eye review stage, the content leads ensured that each item:

- addressed the assigned standard and benchmark, cognitive complexity level, and difficulty level;
- was grade appropriate;
- provided only one correct answer;
- did not contain information in the stem that cued the correct answer;
- incorporated elements of Universal Design and was free of bias or sensitivity issues;
- was consistent with MCESA style guidelines; and
- did not contain any content errors.

WestEd's editing process requires that the content lead is the first editor (as intake editor) and the last editor (as final-eye reviewer) to review each item. This provides a high degree of quality control in the item-development process. The content lead's vision of the full set of developed items is based on the item order and its accompanying information (targeted standards, cognitive complexity, and difficulty). By reviewing the items at both ends of the development process, the content lead ensures that the final set serves its intended purpose of contributing to a robust item pool.

The item development process described in this section has been designed to achieve the quality assurance that MCESA expected. The training procedures described in this section are a critical component of WestEd's approach to quality assurance. While effective training was one pillar of high quality, rigorous processes and systems were also integral pillars to WestEd's approach. WestEd staff shared a commitment to quality and a sense of partnership with MCESA in the product. To that purpose, redundancies were built into WestEd's different reviews, to ensure the integrity of the items' content, appearance, and features of style.

### *4.1.7 Bias/Sensitivity and Item Content Reviews*

Bias review sessions for all assessment items were conducted by Maricopa County non-teachers with no content knowledge. Item content review sessions for all assessment items were conducted by MCESA educators.

The purposes of the bias reviews were to verify that the items were free of stereotypes or other sources of bias and to confirm that they reflected community standards. Participants were asked to ensure that the content of each item was free of explicit references to or descriptions of events involving extreme sadness or adversity, acts of physical or psychological violence, alcohol or drug abuse, vulgar language, or sex. Throughout the bias review, participants were asked to ensure that more than one point of view was expressed when any religious, political, social, or philosophical issues were addressed; that beliefs or biases did not interfere with factual accuracy; that contemporary issues that had already been proven to be controversial were absent from the items; and that stereotypical descriptions of beliefs or customs were absent from the items.

Participants in the bias/sensitivity were selected on the basis of their ability to ensure ethnic, racial, and gender representation.

At the beginning of the bias reviews, the terms "validity," "bias," "sensitivity," and "fairness" were defined. Bias review participants received training in how to ensure that each item:

- was free of offensive, disturbing, or inappropriate language or content;
- was free of stereotyping based on gender, race, ethnicity, religion, socioeconomic status, age, regional or geographic area, disability, or occupation;
- demonstrated sensitivity to historical representation of groups; and
- was free of differential familiarity for any group, based on language, socioeconomic status, regional or geographic area, or prior knowledge or experiences unrelated to the subject matter being tested.

The purposes of the content reviews were to verify the accuracy, difficulty range, depth of knowledge, and grade appropriateness of potential test items and to verify their alignment to the intended Arizona POs.

Reviewers participating in the content reviews received training in how to ensure that the content of each item:

- was targeted to assess only one PO;
- dealt with material that was important in testing the targeted PO;
- was grade appropriate;
- used appropriate thinking skills (e.g., application, analysis, conclusions, extending);
- was presented at a reading level suitable for the grade level being tested;
- was accurate and documented against reliable, up-to-date sources;
- included a stem that facilitated answering the question or completing the statement without looking at the answer choices;
- included a stem that did not present clues to the correct answer choice;
- included answer choices that were plausible and attractive to students who had not mastered the objective or skill;
- was conceptually, grammatically, and syntactically consistent both between the stem and the answer choices and among the answer choices;
- included mutually exclusive distracters; and
- included one and only one correct answer choice.

During the bias/sensitivity and item content reviews, participants were frequently encouraged to discuss each item and to make revisions that would bring the item into compliance with the

preceding conditions. Bias and item content reviews were conducted separately for each item; however, each item earned just one combined rating: "Accept As Is," "Accept with Revisions," or "Reject." Explanations of why an item was rejected were required.

Overall, the item acceptance rates were high. The rates for Phases I and II were 94.36% for the post-assessment and 96.86% for the pre-assessment. For Phase III, the post-assessment rate was 97.54% and the pre-assessment rate was 94.58%. These acceptance rates include reviewed items that were accepted either as is or with revisions. Any item that was rejected, for either bias or content issues, was removed from consideration for field testing. Tables 36 - 39 show the numbers and percentages of items classified into each category during bias and item content reviews, by grade and subject.

**Table 36 – Phases I and II Post-Assessment Bias and Item Content Review Results**

| Grade/Subject | Items Reviewed | Accepted | Accepted % | Rejected | Rejected % |
|---|---|---|---|---|---|
| PE Grade 3 | 167 | 159 | 95.21% | 8 | 4.79% |
| PE Grade 8 | 156 | 150 | 96.15% | 6 | 3.85% |
| PE HS | 166 | 161 | 96.99% | 5 | 3.01% |
| Dance | 164 | 152 | 92.68% | 12 | 7.32% |
| Music Grade 3 | 162 | 143 | 88.27% | 19 | 11.73% |
| Band | 164 | 150 | 91.46% | 14 | 8.54% |
| Choir | 148 | 147 | 99.32% | 1 | 0.68% |
| Theater | 99 | 86 | 86.87% | 13 | 13.13% |
| Theater PB | 69 | 54 | 78.26% | 15 | 21.74% |
| Visual Arts G3 | 143 | 142 | 99.30% | 1 | 0.70% |
| Visual Arts G8 | 141 | 139 | 98.58% | 2 | 1.42% |
| Visual Arts HS | 160 | 158 | 98.75% | 2 | 1.27% |
| Total | 1739 | 1641 | 94.36% | 98 | 5.64% |

**Table 37 – Phases I and II Pre-Assessment Bias and Item Content Review Results**

| Grade/Subject | Items Reviewed | Accepted | Accepted % | Rejected | Rejected % |
|---|---|---|---|---|---|
| PE Grade 3 | 174 | 171 | 98.28 | 3 | 1.72 |
| PE Grade 8 | 174 | 170 | 97.70 | 4 | 2.30 |
| PE HS | 179 | 173 | 96.65 | 6 | 3.35 |
| Dance | 181 | 171 | 94.48 | 10 | 5.52 |
| Music Grade 3 | 166 | 162 | 97.59 | 4 | 2.41 |
| Banks | 172 | 166 | 96.51 | 6 | 3.49 |
| Choir | 176 | 175 | 99.43 | 1 | 0.57 |
| Theater | 176 | 167 | 94.89 | 9 | 5.11 |
| Visual Arts Grade 3 | 165 | 156 | 94.55 | 9 | 5.45 |
| Visual Arts Grade 8 | 176 | 171 | 97.16 | 5 | 2.84 |
| Visual Arts HS | 172 | 169 | 98.26 | 3 | 1.74 |
| Total | 1911 | 1851 | 96.86 | 60 | 3.14 |

**Table 38 – Phase III Pre-assessment Bias and Item Content Review Results**

| Pre- Assessment | | | | | |
|---|---|---|---|---|---|
| Grade/Subject | Items Reviewed | Accepted | Accepted % | Rejected | Rejected % |
| Social Studies Grade 6 | 138 | 129 | 93.48 | 9 | 6.52 |
| Social Studies Grade 7 | 174 | 168 | 96.55 | 6 | 3.45 |
| Social Studies Grade 8 | 187 | 180 | 96.26 | 7 | 3.74 |
| Economics | 126 | 121 | 96.03 | 5 | 3.97 |
| Civics/Government | 112 | 106 | 94.64 | 6 | 5.36 |
| World History | 127 | 110 | 86.61 | 17 | 13.39 |
| U.S. History | 95 | 93 | 97.89 | 2 | 2.11 |
| Total | 959 | 907 | 94.58 | 52 | 5.42 |

**Table 39 – Phase III Post-assessment Bias and Item Content Review Results**

| Post-Assessment | | | | | |
|---|---|---|---|---|---|
| Grade/Subject | Items Reviewed | Accepted | Accepted % | Rejected | Rejected % |
| Social Studies Grade 6 | 216 | 208 | 96.30 | 8 | 3.70 |
| Social Studies Grade 7 | 212 | 210 | 99.06 | 2 | 0.94 |
| Social Studies Grade 8 | 213 | 209 | 98.12 | 4 | 1.88 |
| Economics | 215 | 215 | 100.0 | 0 | 0.00 |
| Civics/Government | 212 | 204 | 96.23 | 8 | 3.77 |
| World History | 218 | 204 | 98.62 | 3 | 1.38 |
| U.S. History | 219 | 218 | 99.54 | 1 | 0.46 |
| Total | 1505 | 1468 | 97.54 | 26 | 2.46 |

## 4.2 Field Testing

### *4.2.1 Development of Pre- and Post-Assessment Field Tests*

Field-test item selection was performed by WestEd for the Phases I and II pre- and post-assessments and by MCESA for the Phase III pre- and post-assessments. In order to ultimately contribute to an item bank of items that measure and support the curriculum and state content standards, selection of the field-test items was guided by the test blueprints.

Items were field tested during four designated pre- and post-assessment windows of time. The following testing windows were used:

- Phase I post-assessment: May 7, 2012, to May 18, 2012
- Phase I and II pre-assessment: October 29, 2012, to November 16, 2012
- Phase II and III post-assessment: May 1-31, 2013
- Phase III pre-assessment: September 3-21, 2013

### *4.2.2 Phase I Post-Assessment Field Test*

Each MCESA grade and subject in Phase I was field tested with three comparable post-assessment forms (A, B, and C) for a total of 107 field-tested items per grade and subject.

All items per form matched the blueprint percentages of items per grade or subject. All items on each form included a balance of easy, medium, and hard items as designated by the item content review committees. Items rated at all difficulty ratings were spread throughout the forms; easy items were not placed only at the beginning of each form, nor were hard items placed only at the end of each form. Items that cued each other were placed on separate forms. Answer keys were checked for each form, in order to ensure that there were not more than three of the same answer choices (A, B, C, or D) in a row. Correct answer choices were balanced across the forms. Each form had a total of 45 items: 31 unique items per form and 14 linking items shared across forms. The linking items matched the blueprint percentages of linking items per grade or subject. The locations of the linking items were exactly the same across all three forms per content area. The linking items included easy, medium, and hard items, as rated by the item content committees.

The use of the shared linking items made it possible to conduct item response theory (IRT) analysis on overlapping data sets in order to establish a common scale for the parameter estimates generated for the alternate forms for each grade and content area. One-parameter and three-parameter (Rasch) models were used to analyze the individual assessment forms. The results of these analyses can be found in Item Response Theory Analysis of MCESA Spring, 2012 Field Tests by Christine G. Burnham, Ph.D., and Jason K. Feld, Ph.D., from Assessment Technology Incorporated (ATI), copyrighted 2012.

The spring 2012 Phase I field tests were administered in 12 school districts located within Maricopa County. Three districts had free-and-reduced lunch count averages at 60% or below. The other nine districts had free-and-reduced lunch count averages ranging from 61-97%. Thus the socio-economic demographic of the field-test participants was skewed slightly from the

Arizona average of 57% free-and-reduced lunch. The field tests were administered for choir, grade 3 music, physical education grades 3 and 8, and visual arts grades 3 and 8. Table 40 presents the numbers of students who took each field-test assessment form.

**Table 40 – Numbers of Students Scored in the Spring 2012 Phase I Field Tests**

| Subject and Field-test Form | Number of Students Scored |
|---|---|
| Choir Form A | 107 |
| Choir Form B | 95 |
| Choir Form C | 58 |
| Music Grade 3 Form A | 848 |
| Music Grade 3 Form B | 894 |
| Music Grade 3 Form C | 855 |
| Physical Education Grade 3 Form A | 871 |
| Physical Education Grade 3 Form B | 951 |
| Physical Education Grade 3 Form C | 911 |
| Physical Education Grade 8 Form A | 672 |
| Physical Education Grade 8 Form B | 546 |
| Physical Education Grade 8 Form C | 682 |
| Visual Arts Grade 3 Form A | 972 |
| Visual Arts Grade 3 Form B | 1012 |
| Visual Arts Grade 3 Form C | 973 |
| Visual Arts Grade 8 Form A | 514 |
| Visual Arts Grade 8 Form B | 454 |
| Visual Arts Grade 8 Form C | 513 |

### *4.2.3 Phases I and II Pre-Assessment Field Testing*

Each MCESA grade or subject in Phases I and II was field tested with three comparable pre-assessment forms (A, B, and C). Each form had a total of 50 items, 17 of which were linking items shared across the forms.

Of the 50 items per form, there were 33 unique items and 17 linking items shared across the forms. Of the 33 unique items, 13 (40%) were entry-level items and 20 (60%) were summative-level items. Of the 17 linking items, seven (40%) were entry-level items and 10 (60%) were summative-level items. The grade 3 physical education entry-level items were written to the K–2 physical education POs, and the grade 3 music entry-level items were written to the grade 2 music POs. The remaining subjects in Phases I and II were written to the post-assessment POs that were designated as entry-level skills by the item specifications committees. The entry-level items were placed more toward the beginning of each form. The locations of the linking items were exactly the same across all three forms.

The theater pre-assessment did not follow the placement of items across forms, due to the excerpts and related items. One entry-level excerpt with related items was placed in the first third of the assessment; one summative excerpt with attached items was placed in the second third of the assessment; and one summative excerpt with attached items was placed in the final third of the assessment.

The 10 summative-level linking items acted as anchor items between the pre-assessment and the post-assessment. For Phase I, summative field-tested items from spring 2012 were used for these anchor items, and were placed in positions relative to their placement on the field tests. These items had a one-parameter range of item difficulty from -1.0 to 1.0. Because the Phase II summative items had not been field tested and did not have statistics, these items were chosen using best judgment of item quality and difficulty. The summative anchor items were tested evenly across each form.

Each form that included unique items and linking/anchor items matched the test blueprint percentages for each item type for each grade and subject. Each grade and subject had an entry-level blueprint and a summative blueprint. All items on each form included a balance of easy, medium, and hard items as designated by the item content review committees. Items rated at all difficulty ratings were spread throughout the forms; easy items were not placed only at the beginning of each form, nor were hard items placed only at the end of each form. However, as previously noted, entry-level items were placed more toward the beginning of each form. Items that cued each other were placed on separate forms. Answer keys were checked for each form, in order to ensure that there were not more than three of the same answer choice (A, B, C, or D) in a row. Correct answer choices were balanced across the forms.

The fall 2012 Phase I and II field tests for pre-assessment forms were administered in 14 school districts located within Maricopa County. Four districts had free-and-reduced lunch count averages at 60% or below. The other 10 districts had free-and-reduced lunch count averages ranging from 61-97%. Thus the socio-economic demographic of the field-test participants was skewed slightly from the Arizona average of 57% free-and-reduced lunch. The field tests were administered in 11 different fine arts and physical education courses. Table 41 presents the numbers of students who took each pre-assessment field-test form.

**Table 41 – Numbers of Students Scored in the Fall 2012 Phases I and II Field Tests**

| Subject and Field-test Form | Number of Students Scored |
|---|---|
| Choir Form A | 200 |
| Choir Form B | 45 |
| Choir Form C | 99 |
| Music Grade 3 Form A | 1349 |
| Music Grade 3 Form B | 1456 |
| Music Grade 3 Form C | 747 |
| Physical Education Grade 3 Form A | 1574 |
| Physical Education Grade 3 Form B | 989 |
| Physical Education Grade 3 Form C | 1586 |
| Physical Education Grade 8 Form A | 1407 |
| Physical Education Grade 8 Form B | 804 |
| Physical Education Grade 8 Form C | 547 |
| Visual Arts Grade 3 Form A | 1971 |
| Visual Arts Grade 3 Form B | 1035 |
| Visual Arts Grade 3 Form C | 837 |
| Visual Arts Grade 8 Form A | 640 |
| Visual Arts Grade 8 Form B | 995 |
| Visual Arts Grade 8 Form C | 511 |
| Band Form A | 340 |
| Band Form B | 214 |
| Band Form C | 76 |
| Theater Form A | 420 |
| Theater Form B | 90 |
| Theater Form c | 43 |
| Dance Form A | 488 |
| Dance Form B | 118 |
| Dance Form C | 87 |
| HS PE Form A | 1102 |
| HS PE Form B | 252 |
| HS PE Form C | 264 |
| HS Visual Arts Form A | 505 |
| HS Visual Arts Form B | 129 |
| HS Visual Arts Form C | 137 |

### *4.2.4 Phases II and III Post-Assessment Field Testing*

Phase II field testing of five high school courses occurred in May of 2013 along with the field testing of seven social studies post-assessments from Phase III. Each post-assessment was field tested with three comparable post-assessment forms (A, B, and C). Each form had a total of 50 items: 36 unique items per form and 14 linking items shared across forms. The linking items matched the blueprint percentages of linking items per grade or subject. The locations of the linking items were exactly the same across all three forms; these locations are shown in Figure

4.3.3. The linking items included easy, medium, and hard items, as rated by the item content committees.

All items per form matched the blueprint percentages of items per grade or subject. All items on each form included a balance of easy, medium, and hard items as designated by the item content review committees. Items rated at all difficulty ratings were spread throughout the forms; easy items were not placed only at the beginning of each form, nor were hard items placed only at the end of each form. Items that cued each other were placed on separate forms. Answer keys were checked for each form, in order to ensure that there were not more than three of the same answer choices (A, B, C, or D) in a row. Correct answer choices were balanced across the forms.

The design of 50 items per form in Phase III was selected to enable the vast breadth of social studies content to be assessed but also retain a form that would be able to be administered in one class period. Each social studies assessment included a reading passage to which five questions assessed the Arizona College and Career Ready Standards for Literacy in History and Social Studies. These passages were placed in the middle of the assessment. For each content area the placement was identical across field-test forms. The reading passages and their related items were placed near the center of the assessment in order to balance the required cognitive load of the task with the overall cognitive load of the entire assessment. The central placement of the passage allowed students to engage the content material via multiple choice assessment items, which are generally less cognitively demanding than reading passages. The passage, also, was not placed at the end of the assessment in order to ensure that this difficult task was not the final task required of students.

The spring 2013 Phase II and III field tests for post-assessment forms were administered in 13 school districts located within Maricopa County. Two districts had free-and-reduced lunch count averages at 60% or below. The other 11 districts had free-and-reduced lunch count averages ranging from 61-97%. Thus the socio-economic demographic of the field-test participants was skewed slightly from the Arizona average of 57% free and reduced lunch. The field tests were administered in 12 different fine arts, physical education and social studies courses. Table 42 presents the numbers of students who took each post-assessment field-test form.

**Table 42 – Numbers of Students Scored in the Spring 2013 Phases II and III Field Tests**

| Subject and Field-test Form | Number of Students Scored |
|---|---|
| Band Form A | 121 |
| Band Form B | 142 |
| Band Form C | 109 |
| Theater Form A | 181 |
| Theater Form B | 130 |
| Theater Form C | 126 |
| Dance Form A | 154 |
| Dance Form B | 275 |
| Dance Form C | 104 |
| HS PE Form A | 294 |
| HS PE Form B | 345 |
| HS PE Form C | 251 |
| HS Visual Arts Form A | 206 |
| HS Visual Arts Form B | 338 |
| HS Visual Arts Form C | 305 |
| Social Studies Grade 6 Form A | 979 |
| Social Studies Grade 6 Form B | 1046 |
| Social Studies Grade 6 Form C | 965 |
| Social Studies Grade 7 Form A | 928 |
| Social Studies Grade 7 Form B | 1048 |
| Social Studies Grade 7 Form C | 963 |
| Social Studies Grade 8 Form A | 956 |
| Social Studies Grade 8 Form B | 996 |
| Social Studies Grade 8 Form C | 930 |
| US History Form A | 186 |
| US History Form B | 188 |
| US History Form C | 135 |
| World History Form A | 280 |
| World History Form B | 243 |
| World History Form C | 186 |
| Economics Form A | 110 |
| Economics Form B | 123 |
| Economics Form C | 101 |
| Government Form A | 220 |
| Government Form B | 145 |
| Government Form C | 25 |

### *4.2.5 Phase III Pre-Assessments Field Testing*

Each MCESA grade and subject in Phase III was field tested with three comparable post-assessment forms (A, B, and C). Each form had a total of 50 items: 36 unique items per form and 14 linking items shared across forms. The linking items matched the blueprint percentages of linking items per grade or subject. The locations of the linking items were exactly the same across all three forms. The linking items included easy, medium, and hard items, as rated by the item content committees.

The design of 50 items per form was selected to enable the vast breadth of social studies content to be assessed but also retain a form that would be able to be administered in one class period.

All items per form matched the blueprint percentages of items per grade or subject. All items on each form included a balance of easy, medium, and hard items as designated by the item content review committees. Items rated at all difficulty ratings were spread throughout the forms; easy items were not placed only at the beginning of each form, nor were hard items placed only at the end of each form. Items that cued each other were placed on separate forms. Answer keys were checked for each form, in order to ensure that there were not more than three of the same answer choices (A, B, C, or D) in a row. Correct answer choices were balanced across the forms.

The fall 2013 Phase III field tests for social studies pre-assessment forms were administered in five school districts located within Maricopa County. Three districts had free-and-reduced lunch count averages at 60% or below. The other two districts had free-and-reduced lunch count averages ranging from 61-97%. Thus the socio-economic demographic of the field-test participants was demographically aligned to the Arizona average of 57% free-and-reduced lunch sub-population. The field tests were administered in 7 different social studies courses. Table 43 presents the numbers of students who took each post-assessment field-test form.

**Table 43 – Numbers of Students Scored in the Fall 2013 Phase III Field Tests**

| Subject and Field-test Form | Number of Students Scored |
|---|---|
| Social Studies Grade 6 Form A | 487 |
| Social Studies Grade 6 Form B | 439 |
| Social Studies Grade 6 Form C | 480 |
| Social Studies Grade 7 Form A | 361 |
| Social Studies Grade 7 Form B | 421 |
| Social Studies Grade 7 Form C | 342 |
| Social Studies Grade 8 Form A | 305 |
| Social Studies Grade 8 Form B | 476 |
| Social Studies Grade 8 Form C | 283 |
| US History Form A | 1124 |
| US History Form B | 1043 |
| US History Form C | 1109 |
| World History Form A | 1611 |
| World History Form B | 971 |
| World History Form C | 785 |
| Economics Form A | 735 |
| Economics Form B | 467 |
| Economics Form C | 758 |
| Government Form A | 442 |
| Government Form B | 753 |
| Government Form C | 342 |

## Part 5. Validity Summary

Validity is determined by a compilation of evidence that can support the ability to make inferences about the resulting test data. To establish strong content validity, MCESA completed several actions. All pre- and post-assessment items were written to the Arizona content standards. Committees of Maricopa County educators were brought together to designate which Arizona performance objectives could be assessed with multiple-choice items and which could be assessed with performance-based items. These educator committees also developed blueprints for their content. The blueprints designated the percentages of emphasis across the Arizona strands and concepts.

To strengthen the validity of items, a detailed process was used to guide item development. The educator committees developed item specifications that included specific instructions on what could be assessed in each performance objective. WestEd and MCESA content specialists reviewed these item specifications and validated the committees' work. The item specifications were then used by Maricopa County educators to write the items. The items were extensively reviewed and edited by WestEd and/or MCESA staff, who ensured that the content of the items aligned to the POs, was grade appropriate, met the highest standards for quality. To further validate the items, content committees of Maricopa County educators reviewed the items to ensure that each item measured the assigned PO, was grade specific, and used grade-appropriate language. When needed, items were edited to meet the approval of the committee. Items that did not meet the specifications required for an acceptable item were rejected. Another committee of Maricopa County citizens (non-educators) validated the items as being free of bias and sensitivity issues. If such issues were found, recommendations for change were made to the content committees.

Final validation of the items came with the actual field testing of the items. Items deemed too difficult or too easy, or that were indicated to be biased, were not used as operational items. Items that had very low or negative discrimination metrics were not used operationally. Final operational pre- and post-assessments were constructed utilizing linking items and strict adherence to map blueprints, in the same way that field-test forms were created. Additionally, the final forms required particular attention to detail regarding item cueing throughout the test form, and a balance of answer choices distributed roughly evenly between A, B, C, and D. The entire process of assessment development was conducted at such a rigorous level as to enhance the construct validity as much as possible.

## Part 6. Reliability Summary

Reliability is the consistency of results. MCESA Content Specific assessments are designed and delivered to yield reliable and consistent results. First, they are developed as fixed forms so that all users take the same test with items in a fixed order. Second, scores are calculated by objective machine-scoring, and third each test is accompanied by a Test Administration Manual to provide a standardized administration experience for students and, thus, reduce sources of systematic error.

To monitor the reliability of MCESA Content Specific Assessments, a reliability coefficient is calculated both on field and operational test forms. Table 44 shows the KR-20 values for MCESA Custom Assessment operational test forms that were administered in 2013-2014 in 10 school districts in Maricopa County through MCESA's REIL Grant.

**Table 44 – KR20 Coefficients**

| Course | Pre or Post | # of students | KR20 |
|---|---|---|---|
| 3rd music | pre | 2611 | 0.68 |
| | post | 2646 | 0.75 |
| 3rd art | pre | 3193 | 0.89 |
| | post | 3428 | 0.84 |
| 3rd PE | pre | 3308 | 0.87 |
| | post | 3357 | 0.89 |
| 8th PE | pre | 2460 | 0.84 |
| | post | 2514 | 0.88 |
| 8th art | pre | 1717 | 0.86 |
| | post | 1704 | 0.87 |
| Band | pre | 724 | 0.88 |
| | post | 875 | 0.85 |
| Choir | pre | 543 | 0.78 |
| | post | 494 | 0.72 |
| HS PE | pre | 40 | 0.67 |
| | post | 39 | 0.75 |
| 6th SS | pre | 1891 | 0.58 |
| | post | 1652 | 0.62 |
| 7th SS | pre | 3099 | 0.67 |
| | post | 2967 | 0.64 |
| 8th SS | pre | 3045 | 0.59 |
| | post | 2973 | 0.62 |
| US History | pre | 33 | 0.67 |
| | post | 29 | 0.46 |
| World History | pre | 25 | 0.52 |
| | post | 23 | 0.22 |
| Government | pre | 25 | 0.81 |
| | post | 0 | 0 |
| Economics | pre | 24 | 0.54 |
| | post | 22 | 0.65 |

# Part 7. Scaling and Equating Analyses

American Institutes for Research (AIR) supported psychometric scaling and validation work with MCESA in its development and implementation of 18 content-area assessments. These assessments were developed to support local education agencies (LEAs) for measuring changes in student achievement and supporting evaluation systems. Psychometric scaling and validation supports these uses by both providing quality assurance that scores produced by these assessments are reliable and appropriate for their use, and generating scores that can be used to measure student gains and other inferential analysis.

## 7.1 Methods

The general process used to develop the raw to scale score conversion tables for each of the different content area assessments included the creation of a set of equated item difficulties across the six field-test forms for each content area (i.e., an item bank), and then pre-equating the fall operational pre-assessment forms (which had not yet been administered when equating procedures began). This method allowed for the pre-equating of additional operational test forms comprised of items from the equated item bank, yielding raw score to scale score tables with scale scores that are comparable to those produced from the previously developed and administered fall 2013 operational pre-assessments.

### *7.1.1 Sample and Administration Details*

For each of the 18 different content areas, a set of three pre-assessment field-test forms (A, B, and C) and a set of three post-assessment field-test forms (A, B, and C) were developed and administered to create a pool of items that could be used to populate operational test forms starting in the 2013-14 school year.

Field testing of Phase I and Phase II pre-assessments occurred in the fall of 2013, and field testing of the post-assessments occurred in spring of 2012 for Phase I and 2013 for Phase II.

The administration details for the Phase III social studies assessments followed a different pattern, yet stilled maintained three pre-assessment field-test forms and three post-assessment field-test forms for each of the content areas. The Phase III post-assessment field-test forms were administered in spring 2013, and pre-assessment field-test forms were administered in fall 2013.

Due to variations in the field testing schedule and assessment delivery vendor during the development of the assessments, the linking design (described in greater detail below) was different for some of the assessments. For all content areas, the set of three pre-assessment forms contained a set of common linking items that allowed for equating of the pre-assessment forms to each other. The same was true for the post-assessment field-test forms for all content areas. For 13 content areas, the pre-assessment and post-assessment field-test forms also contained common items that could be used to link the two sets of forms over time.

For the five remaining content areas, the two sets of items were linked together through post-equating after the administration of the fall operational pre-assessment, which contained both pre-assessment and post-assessment items.

### *7.1.2 The Rasch Model*

AIR conducted IRT analysis with the Rasch dichotomous model as implemented by WINSTEPS (Linacre, 2005). WINSTEPS is well established in the testing industry and can be used to fit a number of different Rasch models (i.e., the Rasch dichotomous model, Rasch rating scale model, and Rasch partial credit model). The Rasch (or 1PL) dichotomous model can be written as follows (Rasch, 1960):

$$P(X_{ij} = 1|\theta_i, b_j) = \frac{1}{1 + exp[-(\theta_i - b_j)]}$$

This formula specifies that the probability of a correct response for examinee i on item j is conditional on the difference between the examinee's ability, θi, and the difficulty of the item, bj. The Rasch dichotomous model defines item characteristic curves that are necessarily parallel to one another and do not cross (a property know as invariant item ordering). As a result of this, the raw score is a sufficient statistic for examinee ability, and there is a one-to-one transformation from raw scores to scale scores for a given test form.

As part of the process of calibrating and scaling the data with the Rasch model, AIR produced, evaluated, and compiled a wide range of item statistics (see Appendix B). These item statistics include the following:

Item Fit Statistics. The Rasch model (as implemented by WINSTEPS) produces two separate fit statistics, the "infit" and "outfit" statistics. Fit statistics assess the degree to which the observed responses agree with the expected responses predicted by the model (Smith, Schumacker, & Bush, 1998). Extreme fit statistics imply that the data may contain unsystematic variability and may not conform to model expectations. The outfit statistic is an unweighted statistic that identifies the prevalence of noisy item responses without accounting for the targeting of items to the examinees. The infit is an information weighted fit statistic that upweights the responses of individuals for whom the items are well targeted. Therefore, the outfit tends to be more sensitive to misfit for individuals who are at the tails of the ability distribution, whereas infit is more sensitive to misfit for individuals near the center of the ability distribution. Both statistics are mean square statistics and have an expected value of 1. Values less than 1 indicate that responses were overly consistent with model expectations. Values greater than 1 indicate that responses were less consistent than expected.

Item Difficulty Estimates and SEs. The Rasch model estimates the difficulty of each item in logits (logarithmic units). Generally speaking, values below zero indicate easier items (high probability of correct response), and values above zero indicate difficult items (lower probability of correct response). Estimating these difficulties is important not only for scaling, but also for assessing the content validity of the test forms. The empirical ordering of the items

in terms of difficulty should make conceptual sense given the definition of the underlying construct (e.g., reading comprehension). Furthermore, estimating the item difficulties allows one to examine both the distribution of item difficulty and the targeting of the items to examinee ability. Distribution of item difficulties is particularly important for examining the ability of the test to be responsive to changes in student abilities over time (i.e., protecting against ceiling and floor effects). Each estimate of item difficulty will also include a standard error for that estimate. These are produced automatically by the modeling software.

Point-Measure Correlation. To examine the internal consistency (i.e., the degree to which each item belongs to a group of all items), the point-measure correlations for the items will be examined. Point-measure correlations assess the degree to which responses to a particular item correlate with the person-ability measures for the individuals providing these responses. Correlations near zero indicate items that may be problematic for inclusion in the test, and negative correlations can indicate mis-keyed items.

### *7.1.3 Equating and Linking Design*

MCESA developed six field-test forms for each of the 18 content areas (three pre-assessment field-test forms and three post-assessment field-test forms). These forms were designed to include items which linked the forms both within administration window (i.e., within pre-assessment administration, among forms A, B, and C), and across administration window (i.e., between pre-assessment and post-assessment administration periods). This design was implemented for 13 content areas. For five content areas, the item test pools were equated after the administration of the 2013 operation test forms, which included items from both the pre-assessment and post-assessment field-test forms. Due to low student counts across these content areas, only band had sufficient student counts to conduct a full analysis. For the remaining four content areas, equating analyses could not be conducted fully immediately after test administration. An additional round of testing was required to increase the number of students taking the test. After this additional round, there was sufficient student N in order to equate the item banks of these four content areas and to pre-equate raw score to scale score tables.

Figure 1 shows the linking design for the 13 content areas which had both within administration and across administration linking items. This linking design was the original MCESA plan for creating an equated item bank.

**Figure 1 – Pre-assessment/Post-assessment Field-test Linking Design (Planned Design)**

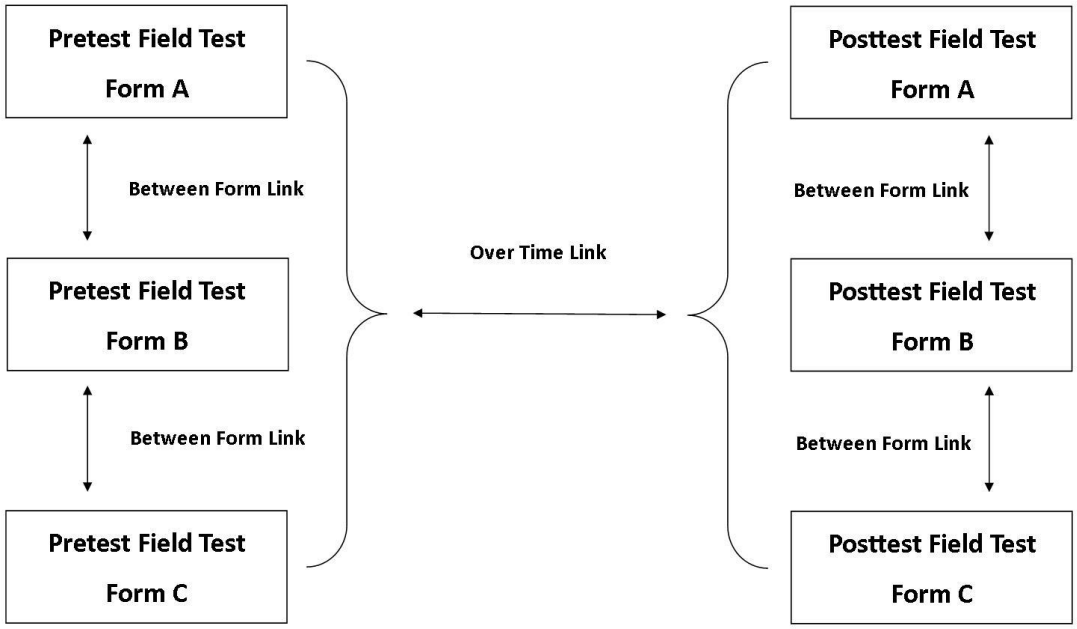
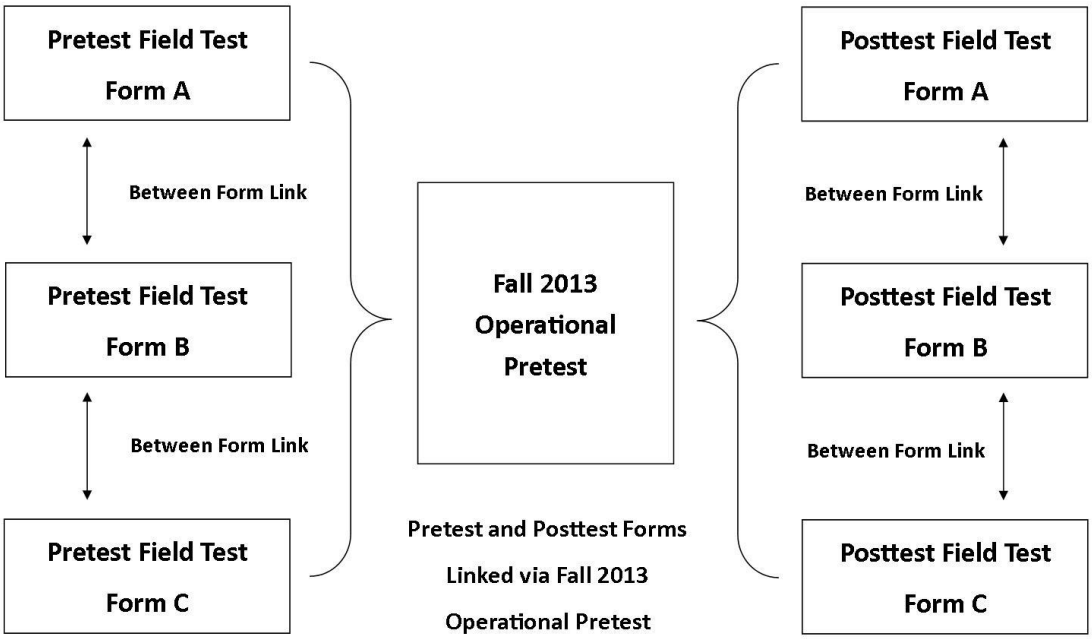


Figure 2 shows the linking design for the five content areas that were equated after administration of the 2013 operational test forms.

**Figure 2 – Pre-assessment/Post-assessment Field-test Linking Design (via Operational Pre-assessment)**




Concurrent calibration using WINSTEPS (Linacre, 2005) was used to equate the item difficulties across field-test forms and administration periods. This method is straightforward and requires only some moderate data manipulation to implement. Data sets from each field test (or operational pre-assessment if applicable) were appended to one another, making sure that data for linking items were matched into the same column in the data set across forms. Items which were unique to a single form were placed into their own unique column. Once the data files were created, the data were calibrated and equated with the Rasch model using the following process:

- Run the stacked concurrent calibration file through WINSTEPS.
- Examine the item statistics for all items. Identify any items with p-values (percent correct) less than .10 or greater than .90, and any items with point measure correlations less than .10.
- Rerun the data set in WINSTEPS, deleting from the analysis any items identified in step two. Output the item difficulties from this WINSTEPS run to be used in step 4.

- Rerun the complete data set (including items identified in step two), anchoring item difficulties for the items calibrated in step 3. This anchors item from pre-assessment and post-assessments to the same metric, based only on well-functioning items. Output the item difficulties from this WINSTEPS run. These item difficulties are the "item bank" difficulties that can be used for pre-equating test forms and developing raw score to scale score conversion tables.

The above process was completed separately for each of the content areas. The process resulted in a set of item difficulties for each of the items administered on the six field-test forms for each content area.

MCESA created fall 2013 operational test forms for each of the content areas consisting of items from the six field-test forms. To expedite reporting of results to teachers, MCESA was interested in providing raw to scale scores for each of the operational pre-assessments prior to completion of pre-assessment administration. Therefore, raw score to scale scores were developed via a process called pre-equating (which uses previously banked values to develop a scale conversion table prior to form administration).

To generate the raw score to Rasch logit score tables, WINSTEPS runs were conducted for each content area using the equated item bank item difficulties for the 50 items on each pre-assessment form. These runs used simulated data and anchored the item difficulties to the item bank values. Scale conversion tables were then output from the software, which included the Rasch logit score associated with each raw score (0 to 50) and a standard error of measurement for that score. The use of simulated data was simply a convenience used to produce the conversion table – the scores depend entirely on the item difficulties of the 50 items comprising the form.

The scale conversion tables produced by WINSTEPS transform raw scores (number correct) into Rasch logit scores. A common scale of 200 to 500 was selected, and so scores were transformed from the logit metric to this score scale. Linear transformations mapping logit scores to this score scale are provided in the results section of this report. It is important to note that additional test forms developed in the future may deviate from the 200 to 500 range depending on the difficulty of the items comprising the new form.

### *7.1.4 Rasch Logit to Scale Score Conversion Formulae*

Table 47 provides the formulae for the linear transformation of Rasch logit scores to scale scores. For each test, the targeted minimum and maximum scores for the fall operational pre-assessment were 200 and 500, respectively. To accomplish this, AIR calculated the equation for the line that mapped the point represented by the lowest logit score and 200 scale units to the point represented by the highest logit score and 500 scale units. Table 45 shows the equations for these lines.

**Table 45 – Rasch Logit to Scale Score Conversion Formulae**

| Test | Scale Conversion Formula |
|---|---|
| Band | Scale Score = 345.9155 + (Logit Score x 28.1690) |
| Choir | Scale Score = 347.9263 + (Logit Score x 27.6498) |
| Dance | Scale Score = 351.9663 + (Logit Score x 28.0899) |
| Music 3 | Scale Score = 354.8928 + (Logit Score x 27.9590) |
| Theater | Scale Score = 348.7443 + (Logit Score x 27.9070) |
| Visual Arts 3 | Scale Score = 349.7207 + (Logit Score x 27.9330) |
| Visual Arts 8 | Scale Score = 351.3838 + (Logit Score x 27.6753) |
| Visual Arts HS | Scale Score = 337.0954 + (Logit Score x 27.7521) |
| Physical Education 3 | Scale Score = 346.7787 + (Logit Score x 28.0112) |
| Physical Education 8 | Scale Score = 349.5890 + (Logit Score x 27.3973) |
| Physical Education HS | Scale Score = 341.6430 + (Logit Score x 28.3286) |
| Social Studies 6 | Scale Score = 346.3483 + (Logit Score x 28.0899) |
| Social Studies 7 | Scale Score = 350.5671 + (Logit Score x 28.3554) |
| Social Studies 8 | Scale Score = 345.6731 + (Logit Score x 28.8462) |
| Government | Scale Score = 343.8863 + (Logit Score x 28.4360) |
| Economics | Scale Score = 343.8318 + (Logit Score x 28.0374) |
| World History | Scale Score = 348.4330 + (Logit Score x 28.4900) |
| US History | Scale Score = 350.7150 + (Logit Score x 28.5987) |

### *7.1.5 Item Bank Size*

Each field-test form (pre-assessments and post-assessments) contained 50 multiple choice items as general practice. However, the final item pool for each content area was not 300 items. Because the forms contained linking items (items common to multiple forms), the number of unique items was lower than 300. Table 46 shows the number of items in the item bank for each of the content areas.

**Table 46 – Item Bank Size by Content Area**

| Test | Total Items | Quality Items | Percent |
|---|---|---|---|
| Band | 232 | 221 | 95.3% |
| Choir | 215 | 191 | 88.8% |
| Dance | 231 | 209 | 90.5% |
| Music 3 | 217 | 202 | 93.1% |
| Theater | 226 | 213 | 94.3% |
| Visual Arts 3 | 219 | 205 | 93.6% |
| Visual Arts 8 | 216 | 207 | 95.8% |
| Visual Arts HS | 233 | 209 | 89.7% |
| Physical Education 3 | 219 | 200 | 91.3% |
| Physical Education 8 | 220 | 206 | 93.6% |
| Physical Education HS | 233 | 217 | 93.1% |
| Social Studies 6 | 205 | 191 | 93.2% |
| Social Studies 7 | 207 | 179 | 86.5% |
| Social Studies 8 | 214 | 198 | 92.5% |
| Economics | 213 | 191 | 89.7% |
| Government | 208 | 187 | 89.9% |
| US History | 205 | 187 | 91.2% |
| World History | 214 | 199 | 93.0% |

## 7.2 Summary

As operational assessment forms were developed for use in 2013-2014 and for future years, IRT item difficulties were used to ensure that raw score to scale score conversion tables for those forms map scores onto a metric comparable with previously developed forms.

## Part 8. Standard Setting

During spring 2015, AIR worked with MCESA to develop and implement a process for setting standards for each of the 18 content-area assessments. MCESA's intention in setting the performance standards for these assessments was to provide information to teachers about their students' overall ability level. Scale scores by themselves provide little information revealing how much content students have mastered. By creating cut scores and reporting categories for students, MCESA can convey additional information to teachers. Because the primary purpose for creating the standards is to provide formative information to teachers on student performance, the stakes associated with these standards are therefore considered to be low. Teachers are not being evaluated on these standards (e.g., based on the percent of students who are "proficient"). Whereas the scale scores that are described elsewhere in this report are being used for growth models, these standards have been created solely to provide information to teachers on student mastery of the material. Because of the low-stakes nature of these standards, MCESA requested that for standard setting AIR develop a statistical approach that relied less on expert judgment than typical high-stakes standard-setting processes do.

MCESA specified that the standard-setting process should establish four category levels. The four score-range categories were labeled as "low," "basic," "medium," and "high":

- Low: Students have demonstrated very low or no understanding of the content associated with the course. Students may need intensive instruction and intervention to meet the course targets.
- Basic: Students have demonstrated limited knowledge and skills associated with this course. They may have many gaps in their knowledge. Students have not yet mastered the course content.
- Medium: Students have demonstrated foundational knowledge and skills associated with this course. While some knowledge gaps might remain, there are fewer than at the basic level. Students are closer to mastering the course content.
- High: Students have demonstrated a strong understanding of the knowledge and skills associated with the course.

The statistical method for setting the cut scores involved determining the scale score at which a student would be expected to answer correctly a certain percent of the item pool. The Rasch dichotomous model (Rasch, 1960) describes the probability of a correct response to a multiple-choice item as a function of the item's difficulty and the ability of the individual responding to the item. This relationship can be used to determine the cut score for a given category.

Each individual test form represents a sample of the content that a student could be expected to know after completing a given course. These items are sampled from the larger bank of possible items that was created during the test-development process. In theory, the item pool represents most of what a student would be expected to learn in the course. However, it is infeasible to ask a student to answer the hundreds of questions in the full item bank. Using the properties of the

Rasch model, AIR determined the estimated percent correct that each student would likely achieve if they were administered the full item bank. This estimated percentage takes into account how each student did on the items on the specific test form, as well as the difficulties of those items relative to the items in the complete item bank. The details of this calculation and how the standards were set are provided below.

Step 1: The initial step would determine the set of items to be considered in the analysis. Each test has an item bank of several hundred items that were written to a specific content standard. First, items that were demonstrated to be too easy (difficulties less than $-2.0$ logits) were removed from consideration. Then, the number of items in the remaining item bank mapped to each content standard was determined and compared to the test blueprint to determine the relative weight of each item. The test blueprint describes how much weight the test gives to each content standard. For standards where there were too many items in the bank, items counted as less than a full item in the analysis. For standards where there were too few items in the bank, items counted as more than a full item in the analysis. This process ensured the representative pool of items used to calculate the estimated percent correct matched the distribution of content described in the test blueprint.

To calculate the weights for each item in the item bank for a specific test, the following formula was used:

$$w_{is} = \left(\frac{N}{n_s}\right) \times BP_s$$

Where $w_{is}$ is the weight for item $i$ mapping to content standard $s$, $N$ is the total number of items in the item bank, $n_s$ is the number of items mapping to the content standard, and $BP_s$ is the percent of items for the standard that are included in the test blueprint. For example, consider the scenario where the item bank contained 100 items, 15 of which mapped to a content standard, but the blueprint indicated that this content area should count for 10 percent of a student's score. In this case, items from that content standard would be given a weight of 0.667.

Step 2: With the weights for each item in the item bank determined in Step 1, the next step involved calculating for each scale score, the estimated percent correct if the student were administered the entire item bank (reweighted to represent the test blueprint). The Rasch model used for the scaling and equating provides an easy way to calculate this percent. Based on a student's scale score and the difficulty of the item, the model calculates a probability of correct response (see Methods: The Rasch Model). Applying the item weights, this calculation is made for all items in the bank, and the probabilities are summed across items to determine the estimated percent correct. The formula for the calculation is as follows:

$$EPC_{ss} = \frac{1}{W}\sum_{s=1}^{S}\sum_{i=1}^{n_s}\frac{w_{is}}{1 + exp[-(\theta_{SS} - b_{is})]}$$

Where $EPC_{ss}$ is the estimated percent correct corresponding to scale score $SS$, $\theta_{SS}$ is the logit ability corresponding to scale score $SS$, $b_{is}$ is the item difficulty for item $i$ from standard $s$, and $W$ is the sum of the item weights.

Step 3: Once the estimated percent correct conversions were calculated for each test, MCESA (with support from AIR) held a series of standard-setting meetings with content-area experts. During these meetings, attendees received information on the psychometric scaling and equating of the assessments, the rationale for setting standards, and the process that would be used to set standards. Experts were then asked to consider (within content-area groups) the percent correct that students in each of the four performance levels would be expected to achieve on average (e.g., 75 percent to be in the high category), based on their knowledge of the courses, their students' abilities, and the item bank's content. This process involved discussion among the content-area experts facilitated by a MCESA staff member, who also provided information on the percent of students who would fall into the four categories based on the current placement of the cut. Following several hours of discussion, the experts indicated their final preferences for the cut scores.

Step 4: Using the cut scores determined by the expert committees, AIR calculated the final cut scores using the formula described in Step 2. Each cut score (represented in estimated percent correct) was mapped backwards to a scale score on the assessment, which then served as the cut score that would be used for the operational assessments.

After cut points had been calculated MCESA prepared communication materials to support teachers and leaders with use of the data.

## Part 9. Conclusion

MCESA has created assessments for 18 special area and non-core content subjects. Every effort has been made to ensure the assessments are valid and reliable for use in educator effectiveness models or other summative uses.

Additional support materials are available for interpreting the MCESA Content Specific Assessments.

- Scale Score and Level Descriptors – This manual provides conversion tables for the raw scores to scale scores and levels as well as the descriptions of each level.
- Score Conversion Tool – This Excel file allows users to insert raw scores from an assessment and the scale score and level will automatically be calculated.
- Test Administration Manual – This manual describes the administration procedures.

## Part 10. References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: Authors.

Burnham, Christine G., Feld, Jason K. (2012). Item Response Theory Analysis of MCESA Spring, 2012 Field Tests. Assessment Technology Incorporated (ATI).

Haladyna, T.M. (1997). *Writing test items to evaluate higher order thinking*. Boston, MA: Allyn and Bacon.

Linacre, J. M. (2005). *WINSTEPS: Rasch measurement computer program* [Computer software]. Chicago: Winsteps.com.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research. (Exp. ed., 1980. Chicago: University of Chicago Press.)

Smith, R. M., Schumacker, R. E., & Bush, M. J. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement*, 2(1), 66–78.

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved [8/31/2015] from the World Wide Web: http://education.umn.edu/NCEO/OnlinePubs/Synthesis44.html